9-2-2016

Bayesian Analysis of Item Response Theory and its Applications to Longitudinal Education Data

ABHISEK SAHA
University of Connecticut, abhiseksaha.isi@gmail.com

Bayesian Analysis of Item Response Theory and its Applications to Longitudinal Education Data

Abhisek Saha, Ph.D.
University of Connecticut, 2016

ABSTRACT

Inferences on ability in item response theory (IRT) have been mainly based on item responses while response time is often ignored. This is a loss of information especially with the advent of computerized tests. Most of the IRT models may not apply to these modern computerized tests as they still suffer from at least one of the three problems, local independence, randomized item and individually varying test dates, due to the flexibility and complex designs of computerized (adaptive) tests. In Chapter 2, we propose a new class of state space models, namely dynamic item responses and response times models (DIR-RT models), which conjointly model response time with time series of dichotomous responses. It aims to improve the accuracy of ability estimation via auxilary information from response time. A simulation study is conducted to ensure correctness of proposed sampling schemes to estimate parameters, whereas an empirical study is conducted using MetaMetrics datasets to demonstrate its implications in practice. In Chapter 3, we have investigated the difficulty in implementing the standard model diagnostic methods while

comparing two popular response time models (i.e., monotone and inverted U-shape). A new variant of conditional deviance information criterion (DIC) is proposed and some simulation studies are conducted to check its performance. The results of model comparison support the inverted U shaped model, as discussed in Chapter 1, which can better capture examinees' behaviors and psychology in exams. The estimates of ability via Dynamic Item Response models (DIR) or DIR-RT model often are non-monotonic and zig-zagged because of irregularly spaced time-points though the inherent mean ability growth process is monotonic and smooth. Also the parametric assumption of ability process may not be always exact. To have more flexible yet smooth and monotonic estimates of ability we propose a semi-parametric dynamic item response model and study the robustness of the proposed model. Finally, as every student's growth is different from others, it may be of importance to identify groups of fast learners from slow learners. The growth curves are clustered into distinct groups based on learning rates. A spline derivative based clustering method is suggested in light of its efficacy on some simulated data in Chapter 5 as part of future works.

Bayesian Analysis of Item Response Theory and its Applications to Longitudinal Education Data

Abhisek Saha

B. Stat., M. Stat., Statistics, Indian Statistical Institute, India, 2007

M.S., Statistics, University of Connecticut, CT, USA, 2015

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Bayesian Analysis of Item Response Theory and its Applications to Longitudinal Education Data

Presented by

Abhisek Saha, B. Stat. Statistics, M.S. Statistics

Co-Major Advisor _____

Dipak K. Dey

Co-Major Advisor _____

Xiaojing Wang

Associate Advisor _____

Ming-Hui Chen

University of Connecticut

2016

Acknowledgements

This dissertation would never have been possible without the support of many individuals, and it is with great pleasure that I use this space to acknowledge them.

I would like to express my utmost gratitude to both of my advisors in general, in particular to Dr. Xiaojing Wang , for being extremely patient with me and helping me debugging codes on many occasions, and to my other advisor, Prof. Dipak Dey for giving me the freedom to explore unfamiliar territory and guiding me through out and supporting me morally. Both taught me courses that motivated to continue research in topics I chose . I am very thankful to Prof. Ming-Hui Chen, for examining my proposal first, eventually examining my thesis and being part of my dissertation committee. I would like to hereby thank Prof. Jun Yan for the class project that greatly helped in deciding what I wanted to work on. Let me hereby thank Jack Stenner, Carl Swartz, Donald Burdick, Hal Burdick and Sean Hanlon at MetaMetrics Inc. for generously sharing the data with us. I want to thank all of the faculty in the statistics department for providing me with a strong technical foundation on which I complete my research and for providing administrative, computing and financial support through Grants. Many fellow graduate students have helped me live though the graduate life. I shall always cherish their friendship deep down. Finally and most importantly, none of this would have been possible without the love and patience of my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Item Response Theory

In psychometrics, Item Response Theory (IRT) is a very popular paradigm that deals with designing, analyzing, and scoring of tests, questionnaires etc that measure abilities, attitudes, or other latent traits. This is the reason why it has another name, latent trait analysis. To understand the estimation process better we consider the analysis of test data consisting of a number of multiple choice type questions. In this test we assume that 100 students are given a mathematical placement test and each test contains 20 multiple choice type questions on topics in college algebra. The intended test in this case is supposed to assess the students mathematical ability and accordingly should help decide which mathematics course would be ideal for him/her.

In designing such a test there are some immediate concerns. The test designer intends to put items in different levels of difficulty. If all items are very easy in relative to general levels of abilities then all students will get them right and the results may not be helpful in accurately assessing their math proficiency. Similar situation is anticipated if the items are too difficult. So generally it is desirable to have broad range of performances

on the exam. This wide difference in performance will help assessing the ability in a better way. This aspect is usually addressed through difficulty parameter which will be elaborated later in equation 1.1.

If the difficulty is properly addressed, there could be another concern. The concern is if the items can discriminate between students. For example, if an item is incorrectly answered by all of the people then it is useless for the estimation. Hence nothing would be lost if it is removed. So an "ideal" item is the one that students having ability below its difficulty level get it incorrect whereas students having ability larger than its difficulty get it correct. Probably such item does not exist in reality but most valuable items are those that exhibit strong positive correlation with math proficiency.This aspect is addressed through discriminatory parameter as will be elaborate in equation 1.1.

To put things in perspective, these test characteristics and abilities can be learned through what is known as item response models. The model represents the probability that a student answers an item accurately. Usually the probability is a function of students ability and other two parameters are item difficulty. This latent score is better than a test score for assessing ability because they let compare scores across many tests with similar purposes. Tests with similar purposes can be very different in terms of design. So test scores may not be comparable. In the next section we discuss some popular IRT models and discuss their properties along with their limitations.

1.2 Rasch Model and its Variants

1.2.1 Rasch Models and its 2 Parameter, 3 Parameter Versions

Suppose θ_i denotes i-th person's ability and $X_{i,l}$ is his/her binary response to the item l (1 if accurate else 0). One of the popular probability models can be given by what is known as 2 parameter IRT models.

$$\Pr(X_{i,l} = 1 \mid \theta_i, d_l, a_l) = \mathrm{F}(d_l(\theta_i - a_l)), \tag{1.1}$$

$$d_l: \text{ discriminatory parameter}; a_l: \text{ difficulty parameter},$$

Here $F(x)$ can be any distribution function. But the most popular choices in the literature are logistic distribution ($F(x) : [1 + e^{-x}]^{-1}$) and normal ogive distribution ($F(x) : \Phi(x)$). Both are quite popular for their own attractive properties. Logit link can be expressed as a log odds ratio while normal ogive may be easier to work with in Bayesian computations. One can note that if items of large a_l values are chosen probability of correct response is going to be very low for all students. This is the difficulty aspect mentioned in earlier section. Here we also note that if a_l and θ are treated fixed probability of correct response will go up if θ is larger than a_l, otherwise it will decline. This is the discriminatory aspect stressed in the last section. When all the items are assumed to have same discriminatory power d_l becomes 1 and this is what Rasch (1961)

proposed along with the choice of $F(x)$ to be logistic distribution.(also called 1 parameter logistic (1-PL)). If $F(x)$ is logistic 1.1 is called 2 parameter logistic model (2-PL). It was then extended to 3 parameter logistic model (3-PL) by incorporating a guessing parameter, that indicates prior knowledge that the item carries about the answer.

$$\Pr(X_{i,l} = 1 \mid \theta_i, d_l, c_l^*, a_l) = c_l^* + (1 - c_l^*)\mathrm{F}(d_l(\theta_i - a_l)) \tag{1.2}$$

c_l^* = guessing parameter.

1.2.2 Implicit Assumptions in Rasch type Models

Traditionally, almost all variants of Rasch models assume *local independence* in IRT, which means, conditionally on θ_i, d_l, a_l as in (1.1) the responses are independent. However, let us consider answering a few multiple choice type questions based on a passage where one usually answers a question once he/she has an overall comprehension of the passage. Clearly here, the local independence assumption falls apart. Such questions are more common in today's tests.

With the advent of computerized tests, the modern test formats have gone through significant changes. Modern test formats differ from the classical ones on the following aspects: (a) while classical test data used to be collected at a single point of time, the new computer based tests allow students to take tests over different times. Note that in classic Rasch model ability is not treated dynamic. In addition, new computerized tests can

be taken at *individually varying time intervals*, which can only add to the complexity. (b) With the advent of computerized tests, the new information about the test-taker background along with *response time* taken now can be recorded. Rasch model did not allow co-lateral or co-variate information. (c) Classical test formats used to present each test taker with the same set of questions (items), thus item-wise calibration was possible based on many samples, while on the other hand computerized tests do not allow for calibration since (i) the tests let students choose passage from a pool of articles, often based on some estimate of his/her current ability. As a result, two students usually do not pick the same passage; (ii) even in case two people choose the same passage they are usually asked randomly selected different subsets of questions from the set of all questions that can be asked from the passage. This phenomenon is often referred to as *randomized item*.

These changes necessitate revising the classical models and adapting them to a set of new assumptions or introducing new set of models to address them entirely. Next we elaborate on these changes and discuss recent developments addressing them.

1.3 Recent Developments in Response Models

1.3.1 Local Dependence and Randomized Item

For local dependence issues, there have been parallel developments in recent years. These works can either be of two types: (1) detecting local dependence through formulating

tests. For example, Chen and Thissen (1997), Glas and Falcón (2003) built χ^2 based test and score tests respectively, whereas Liu and Maydeu-Olivares (2013) worked with a general purpose statistic (called R_2 which is asymptotically equivalent to a χ^2) to determine local dependence. However, some of these tests may be defined on minimal assumptions on information matrix approximation but may compromise on power. (2) Others worked towards modeling these dependencies (Jannarone (1986), Andrich and Kreiner (2010) and Wang, Berger, and Burdick (2013)). For example, Andrich and Kreiner (2010) tried modeling conditionals of consecutive item selection, while Wang et al. (2013) brought in the idea of random test effects and daily effects.

To allow for randomized items (as discussed in subsection 1.2.2), recent works usually bring random effects to model it in IRT. Sinharay, Johnson, and Williamson (2003), De Boeck (2008), Wang et al. (2013) took care of it by adopting this approach.

1.3.2 Longitudinal IRT

In this thesis we have worked with longitudinal data, in which a person can sit for multiple tests at different dates. We are interested to study the growth of the latent trait (ability in our case). So individual's ability is not constant over time. This idea necessitates a growth process of ability over time, which can not be accommodated by traditional models. The recent works seem to approach these issues in one of 3 ways: (a) *by parametric function of time*, for example, Johnson and Raudenbush (2006) modeled ability by linear or polynomial function of time, where these time points are equispaced

and fixed for all test-takers whereas Hsieh, von Eye, Maier, Hsieh, and Chen (2013) came up with couple of inter-dependent structural equations involving time. Verhagen and Fox (2012) considered linear and quadratic functions of time with random coefficients. (b) *by Markov Chain*, for instance, Park (2011) assumed changes in voting preferences are due to age-specific regime changes and modeled it by a Markov process, whereas Bartolucci, Pennoni, and Vittadini (2011) analyzed tests scores by modeling transition probabilities with covariates; (c) by *a combinations of them*, Bollen and Curran (2004) made a comparative study and showed neither of them could be enough to address latent trajectory models, Wang et al. (2013) combined the two ideas.

1.4 Response Time Models

The relation between response and response time has been debated for years. Roskam (1997), Wang and Hanson (2005) suggested models where they considered response as a causal factor in determining accuracy. As for instance one can observe in the following example of Roskam (1997),

$$\Pr(X_{i,l} = 1 \mid \theta_i, a_l) = \mathrm{F}(\theta_i + logR_{i,l} - a_l), \quad \theta_i = \text{“mental speed”}. \tag{1.3}$$

Here interpretation of θ_i is slightly different. With $\exp(\theta_i + logR_{i,l})$ or $\exp(\theta_i)R_{i,l}$, the product represents the total faculty and it is interpreted as product of “mental speed” (θ in exponential scale) and “time” (R). Yet equation 1.3 becomes a variation of Rasch

model. The model received criticism for treating response time known beforehand. Gaviria (2005) proposed to model log-response time given accurate answer and left it unspecified when inaccurate. Thissen (1983) suggested to model the underlying parameter responsible for accuracy. Thissen (1983) model can be given as follows

$$\log R_{i,l} = \mu + \nu_i + \tau_l + \beta L(\theta_i - a_l) + \zeta_{i,l}. \tag{1.4}$$

Here $L(x)$ denotes a linear function, usually with discrimination factor as introduced in 2-PL models. ν_i and τ_l denote what are called "slowness" parameters. It assumes that response time depends on two quantities, one is speed of the test-taker, which is the amount of time that person takes for infinitely easy set of problems and slowness intensity of the question, which dictates the time taken due to the nature of the problem. In recent years joint hierarchical models were introduced to incorporate both accuracy models along with Thissen (1983)'s type of response time models (RTM), based on the idea that response time should be treated as random variable jointly with accuracy. For instance, Ferrando and Lorenzo-Seva (2007) proposed the joint models conditionally on θ and other factors in which they took an RTM as in (1.4) with $L(x)$ as $\sqrt{Linear(x)^2}$ and 2-PL IRT model. On the other hand, Van der Linden, Klein Entink, and Fox (2010) proposed an RTM as in (1.4) with $L(x)$ as $Linear(x)$ conditionally on a person specific latent parameter (τ) and other item specific parameters along with a 3 parameter normal ogive (3-PNO) for IRT model (similar to 3-PL but $F(x)$ is chosen to be probit instead)

conditionally on latent ability (θ), which were called lower level models. At the higher level, they specified the joint distributions of θ and τ, thereby allowing the information to be borrowed.

1.5 Bayesian Estimation of IRT and its Advantages

In modern tests, inference is drawn in presence of quite complex dependency structures and many sources of uncertainty. Traditional frequentists' methodology has approached the problem through various iterative schemes (such as Expectation Maximization algorithm (EM)) in which each iteration step tries to solve less complex sub-problem and eventually combine the results. For example, in standard marginal maximum likelihood (MML) practice, one estimates items (called item calibration), that is, items are estimated assuming ability is missing (Bock and Aitkin (1981)) and then the item parameters are treated known and fixed at their calibrated values when proceeding with inference regarding examinees and sub-populations. This methodology has been the key to successful implementation of IRT methods. However, as complexity of the model increases, application of EM type algorithm becomes less straightforward. Moreover, as mentioned by Robert K. Tsutakawa (1988), it is hard to incorporate uncertainty into the item parameter estimates for calculations of standard errors about ability estimates. In contrast, in Bayesian methods, while computing maximum a posteriori (MAP) or

expected a posteriori (EAP) estimates of ability in a fully Bayesian framework, estimation uncertainty is automatically incorporated into the standard errors of MAP or EAP estimates. As far as computation is considered, implementing MCMC steps (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2013), Gelfand and Smith (1990), Chib and Greenberg (1995)) is usually simpler than computing quadratures (E-step of EM) or derivatives (M-step of EM). The cost of this flexible implementation is usually slower convergence of estimation algorithms. Albert (1992) first popularized the Bayesian application in a data augmentation version(Tanner and Wong (1987)) of 2 parameter normal ogive (2-PNO) models using Gibbs sampling (Gelfand and Smith (1990)), a very popular MCMC techniques while Patz and Junker (1999) extended the work of Albert (1992) to general problems with methods based on Metropolis Hastings within Gibbs (Chib and Greenberg (1995)). Since then these implementations have been adopted extensively in many Bayesian applications of IRT (Verhagen and Fox (2012), Fox and Glas (2001), etc). In this thesis we have used the data augmentation method in a 1-PL model following the work of Wang et al. (2013).

1.6 Motivation

Although there is a huge literature in IRT models, there have not been much work in the paradigm of longitudinal IRT. This is because of advancement in technology and computerization of tests for which it became possible in recent years to track one's

history with same types of tests. We have discussed some interesting deficiencies of current methodologies throughout previous subsections. These serve as motivations for the development new methodologies in this field. Though DIR (Cf. Wang et al. (2013)) models serve as a unified framework that are flexible to address general longitudinal testing framework, they do not incorporate response time. So DIR models do not give any idea about how speed plays a role in determining response. In addition, most IRT models are too simple to be extended to complex scenarios like test takers coming back and sitting for more than one exam on the same day, at irregularly spaced time points. Joint hierarchical response time models do not assume explicit relationships at higher levels. To improve the precision of estimates by incorporating response time via joint modeling of response time and response, that, in addition, establishes speed-accuracy relation in response time model with the an interpretable linkage, works as a motivation for developing dynamic item responses and response times (DIR-RT) models in Chapter 2. Because of complexity of DIR-RT models, it becomes harder to compare different response time models with DIR-RT modeling framework. This necessitates development of model diagnostic measure that can address this gap satisfactory. As a result, new measures are developed for this purpose in Chapter 3. Next we focused on growth models of ability and we observe that certain aspects of growth trajectories, like smoothness, monotonicity were never addressed directly in the modeling framework. In addition, all these growth models of ability are based on some parametric assumptions. The goal to model ability trajectory with minimal assumptions on growth except for

curvature properties like smoothness or monotonicity (which are usually ignored during IRT modeling) has led us to develop semi-parametric dynamic item response model with monotonic and smooth growth curve in Chapter 4. Finally we have addressed partially relatively less-explored and less-stressed aspect of IRT modeling, clustering ability curves. To the best our of knowledge, clustering ability curves were never explored with an objective to identify different groups of students with significantly different learning rates from one another. This has motivated us, to build a distance-based clustering method to cluster students based on differences in learning patterns and, eventually to propose also a model-based alternative as part of future work in Chapter 5.

1.7 Thesis Outline

In Chapter 2 we introduce DIR-RT models and describe its properties. Later we implement and verify efficient parameters recovery through simulation study. Efficacy and usefulness of DIR-RT models compared to DIR models (Wang et al. (2013)) are established through the same simulation study. Eventually we apply the methodology to MetaMetrics's EdSphere data and response time model with I-U shaped linkage is justified based on empirical evidences. Later in that Chapter, implications of the posterior estimates are discussed. In Chapter 3, we propose to compare between two popular response time (RT) models. We next discuss the difficulty in applying the traditional

measures for model selection and a new model diagnostic criterion is presented, that can be used for model selection. We study the goodness of the criterion through simulation study. In Chapter 4, we propose an alternative ability growth process, a smooth and monotonic semi-parametric growth model. We then discuss the impact of regularization to ensure smoothness. Posterior computation is executed based on simulated examples to ensure efficient parameters estimation. Eventually we study the robustness of semi-parametric model in the context of curve fitting for simulated data from DIR models. In Chapter 5 we summarize the findings from all three Chapters and then suggest a spline derivative based clustering technique as well as a model-based alternative to cluster the ability growth curves based on their shapes. This analysis can be useful in practice to help achieve goals of personalized education.

Chapter 2

Bayesian Joint Modeling of Response Times with Dynamic Latent Ability

2.1 Introduction

Item Response Theory (IRT) models, also known as latent trait (analysis) models have been widely used in testing for several decades. They originated from analyzing dichotomous items (Lord (1953) and Rasch (1961)), soon extended to modeling polytomous items (Samejima (1969) and Darrell Bock (1972))). Their applications became diverse from education and psychology to political science, clinical and health studies, marketing and so on. The popularity of IRT models is because of their separability of assessment of the latent traits of examinees (e.g., attitude, proficiency, preferences and other mental/behavior properties) from effectiveness of the test items. One of the most famous IRT models is Rasch model (Rasch (1961)), belonging to one-parameter IRT

models, which is typically specified as

$$\Pr(X_{i,l} = 1 \mid \theta_i, d_l) = \mathrm{F}(\theta_i - d_l), \tag{2.1}$$

where the subscript (i, l) is used to index i-th person and l-th item (or question), $X_{i,l}$ then represents the correctness (1 if correct otherwise 0) of the answer, d_l denotes the level of item difficulty, and $F(x)$ is the link function. For the Rasch model, the link function is chosen to be logistic.

Traditionally, Rasch models and all variants of them, such as two-parameter or three-parameter IRT models, are based on the *local independence* assumption, which means, conditionally on θ_i, d_l, (as in (2.1)), the item responses $X_{i,l}$'s are statistically independent. Classical IRT models are usually applied to data collected for exams in a paper-and-pencil form, where different examinees take the same test at the same time. However, with the advent of computer-based (adaptive) testing, examinees can take series of tests online or in the classroom at anytime as they wish and items are instead randomly drawn from a bank of items. Then, the changes of test formats necessitate revising assumptions of the classic IRT models and introducing new set of models to accommodate changes.

2.1.1 MetaMetrics Testbed and Recent Developments of IRT Models

Our study is motivated from EdSphere dataset provided by MetaMetrics Inc. EdSphere is a personalized literacy learning platform that continuously collects data about student performance and strategic behaviors each time when he/she reads an article. The data was generated during sessions in which a student read an article selected from a large bank of available articles. A session begins like this: a student selects from a generated list of articles having text complexities (measured in other platform of MetaMetrics test design) in a range targeted to his/her current ability estimate. Once the article is chosen, the computer, following a prescribed protocol, randomly selects a sample of the eligible words to be "clozed", that is to be removed and replaced by blanks and presents the article to the student with these words clozed. When a blank is encountered while reading the article, the student clicks it and then the true removed word along with three incorrect options called foils are presented. As with the target word, the foils are selected randomly according to a prescribed protocol. The student selects a word to fill in the blank from the four choices and an immediate feedback is provided in the form of the correct answer. The dichotomous items produced by this procedure are called "Auto-Generated-Cloze" items and are *randomized items*. The key feature of these items is their single usage, which implies even if two students select that same article to read, the sets of target words and foils will be totally different. As a consequence, it is not

feasible to obtain data-based estimates of item parameters (calibration).

The EdSphere dataset consists of 16,949 students who registered over 5 years in EdSphere learning platform testing from a school district in Mississippi. The students were in different grades and entered and left the program at different times between 2007 and 2011. They can take tests on different days and have different time lapses between tests, which means the observations collected are *longitudinal* at individually-varying and irregularly-spaced time points. Of course, a dynamic structure to modeling changes of latent traits is needed. In addition, as mentioned in Wang et al. (2013), in the environment of EdSphere, the factors such as an overall comprehension of the article (an example of test random effects), the person's emotional status (an instance of daily random effects) and others, might undermine the *local independence* assumption of IRT models.

To summarize, the distinctive features, i.e., *randomized items*, *longitudinal observations*, and *local dependence* often appear in the modern computerized (adaptive) testing (not merely MetaMetrics datasets), making the classic IRT models face great challenges. To address these, there have been many developments. To generalize IRT models for *longitudinal data*, some researchers (e.g, Albers, Does, Imbos, and Janssen (1989), Johnson and Raudenbush (2006), and Verhagen and Fox (2012)) used parametric function of time to model changes of latent traits; while others (Martin and Quinn (2002), Park (2011) and etc.) applied a Markov chain model to describe the time-dependence of latent traits. Yet neither of the two ideas would be enough to describe the changes (Bollen

and Curran (2004)). Instead, Wang et al. (2013) modeled the growth of latent traits by combining the two ideas. For *local dependence*, there have been parallel developments for the procedures of detecting it (e.g., Yen (1984), Chen and Thissen (1997) and Liu and Maydeu-Olivares (2013)) and the ways of modeling it (e.g., Jannarone (1986), Bradlow, Wainer, and Wang (1999) and Cai (2010)). For *randomized items*, introducing random effects for item parameters is often used (e.g., Sinharay et al. (2003) and De Boeck (2008)).

The literature that focuses on three features simultaneously is very limited. However, within one unified framework, Wang et al. (2013) developed a new class of state space models, called Dynamic Item Response (DIR) models, to describe the dynamic growth of an individual's latent trait,that account for local dependence and address uncertainty of test items in the testing. In this regard, their work is pioneering but they ignored the usage of the response time information (often easily obtained during computerized tests) to aid the estimates of one's ability.

Thissen (1983) showed that the separate analysis of response accuracy and response time in a test would be misleading. The analysis of Ferrando and Lorenzo-Seva (2007), Van der Linden et al. (2010) and Ranger and Kuhn (2012) further demonstrated that using response times as auxiliary information can both improve the precision and reduce the bias of the estimates of IRT parameters. Therefore, the joint analysis of response times with item responses in a computerized (adaptive) testing will be a significant advancement of DIR models.

2.1.2 Recent Developments for Modeling Response Times in Educational Testing

To model the response time of an item, one way is to treat it as a causal factor for the accuracy of that item (e.g., Roskam (1997) and Wang and Hanson (2005)). Another idea regards response accuracy as a casual factor for the response time (e.g., Gaviria (2005)). However, both ideas have been criticized since the response time and accuracy of a test may not be directly related. Instead, the third way is to jointly model response times and item responses in a hierarchical fashion.

There are two distinct classes of joint modeling, based on different views of the relationship between response accuracy and response times. The first category conceives of a speed-accuracy tradeoff (Luce (1986)) or a variation of that. A popular choice in the stage of modeling response times is Thissen (1983) model, i.e., taking the natural logarithm of response times and modeling that as follows,

$$\log R_{i,l} = \mu + \nu_i + \tau_l + \beta L(\theta_i - d_l) + \zeta_{i,l}, \tag{2.2}$$

where $R_{i,l}$ indicates the time used for l-th question by i-th person, ν_i is the speediness parameter, which takes account the time that person spends for infinitely easy set of problems, τ_l is the slowness intensity of a question, which dictates the time taken due to the nature of the problem, μ is the overall mean, $\zeta_{i,l}$ is the residual, β is a slope and $L(x)$ denotes a linear function mapping how the distance of ability and item difficulty

connect with response times.

There are two popular choices for $L(x)$, one is a monotone mapping (e.g., Thissen (1983) and Gaviria (2005)), reflecting the idea that the larger the distance is, the more time it costs; the other is an inverted-U (I-U) shaped mapping, originating from the findings (e.g., Wang (2006) and Wang and Zhang (2006)) in educational testing that examinees generally spend more time on items that match their ability levels, while spend less time on items either too easy or too hard. Ferrando and Lorenzo-Seva (2007) and Ranger and Kuhn (2012) also employed the inverted U-shape for regressing response times in the analysis of personality and psychology tests. Intuitively, the negative β in front of $L(x)$ for either monotone or inverted U-shaped mapping makes more sense in reality.

The second category (e.g, Van der Linden (2007), Klein Entink (2009) and Loeys, Rosseel, and Baten (2011)) utilizes a hierarchical framework to jointly model response times and accuracy but without specifying explicit relationship between them. Instead, they assigned joint multivariate normal priors to link parameters of the joint models.

However, all existing joint models are centered on one-time exam for testers without considering the features in computerized testing. In this paper, we aim to fill in this gap. Enlightened by DIR models, we will propose the idea of jointly incorporating response times with response accuracy for testing data collected at irregular and individual varying time points.

2.1.3 Preview

In section 2, we will put forward a new class of joint models for IRT models with response times, which will be called dynamic item responses and response times models (DIR-RT models). In the response model we propose I-U shaped linkage. Because of the complexity of the model considered, Bayesian methods and Markov Chain Monte Carlo (MCMC) computational techniques will be employed. Section 3 will present the statistical inference procedures. Section 4 validates Bayesian inference procedure proposed with some simulations and compare the performance of DIR-RT models with respect to DIR models. We illustrate the application of DIR-RT models to MetaMetrics testbed datasets. In section 5, we further provide an empirical justification of the goodness of the fit for DIR-RT with I-U shaped linkage. In Section 6, we point out some significant psychological results from the analysis of MetaMetrics dataset and show the direction for our future studies.

2.2 Joint Models of Dynamic Item Responses and Response Times (DIR-RT)

Clearly to jointly model (2.1) and (2.2), it will maximize the information to infer one's ability θ_i and the item difficulty d_l. Besides, notice earlier that conducting a separate analysis of response accuracy or response time alone will be misleading since timed tests

usually involve accuracy and time spent as two dimensions. Thus, these motivate us to propose a *two-stage* joint model. The first stage has two sub-models to concurrently model the observations of response time and response accuracy with certain sharing parameters, and the second stage introduces a dynamic model to capture changes of latent traits over time. Although our investigation begins with an extension on one-parameter IRT, it would be straightforward to generalize it to two-parameter or three-parameter IRT models.

2.2.1 First Stage: The Observation Equations in DIR-RT models

(2.1) or (2.2) from the current literature are based on an one-time exam for each test taker, a much simpler situations than that of a computerized test. To accommodate the complication, we first expand the labels of notations.

Let $X_{i,t,s,l}$ be the item response to indicate the correctness of the answer of the l-th item in the s-th test on the t-th day given by the i-th person, where $i = 1, \cdots, n$ (number of subjects); $t = 1, \cdots, T_i$ (number of test dates); $s = 1, \cdots, S_{i,t}$ (number of tests in a day); and $l = 1, \cdots, K_{i,t,s}$ (number of items in a test). Likewise, denote the difficulty of the l-th item as $d_{i,t,s,l}$. It is ideal to record the time for each tester spending on a single item, however, in practice, more often the time spent on the entire exam is merely stored for each individual. This is a case for reading comprehension tests in MetaMetric

testbed. Then, in our proposed models, the response time is defined at the test level, i.e., $R_{i,t,s}$, implying the time spent on s-th test for the i-th individual on the t-th day; whereas, our models can be easily revised to cope with the response time stored for each item whenever such data is available.

The label extension illustrates two major features of computerized (adaptive) testing: 1) the rarity of replication of items among different time, tests and test takers; 2) the observations being recorded at individually-varying and irregularly-spaced times points. Here, $X_{i,t,s,l}$'s and $R_{i,t,s}$'s are observed. Usually, the response time is naturally bounded above zero, and a logarithmic transformation of $R_{i,t,s}$ will be taken to remove its skewness in our models.

The Observation Equations of Item Responses

Often in a design of computerized tests, item difficulty, i.e., $d_{i,t,s,l}$, is a randomized parameter, assuming to be randomly drawn from a bank of item with certain ensemble mean. $d_{i,t,s,l}$ then can be modeled as a measurement error model, where $d_{i,t,s,l} = a_{i,t,s} + \epsilon_{i,t,s,l}$ with $a_{i,t,s}$ being an ensemble mean difficulty of items in the s-th test, and $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with σ^2 known according to the test design, $\mathcal{N}(\cdot, \cdot)$ denoting a normal distribution. Similar as Wang et al. (2013) did, we extend classic IRT models to accommodate the complication by modeling the observation equation of item responses as

$$\Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}) = \mathrm{F}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}), \tag{2.3}$$

where $\theta_{i,t}$ represents the i-th person's ability on the day t with assuming one's ability is constant over a given day, $\varphi_{i,t}$ and $\eta_{i,t,s}$ take account of daily and test random effects, respectively, to explain the possible local dependence of item responses. Assume $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$ with its precision unknown and being different for each person. Similarly, let $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$ with $\eta_{i,t} = (\eta_{i,t,1}, \ldots, \eta_{i,t,S_{i,t}})'$ being the vector of test random effects on the day t for the individual i and $\mathbf{I}$ is an $S_{i,t} \times S_{i,t}$ identity matrix. Utilizing precision parameters in place of variance parameters for normal distributions is because of the convenience in Bayesian computation. The reason of letting $\eta_{i,t}$ be a singular multivariate normal (by setting the test random effects to be zero on a day t) is to remove any possibility of unidentifiable issues between daily and test random effects. In the application to MetaMetrics testbed, choose $F(x)$ to be a logistic link due to the convention in MetaMetrics, where they used logit unit as a linear transformation of Lexile scale used in their products.

The Observation Equations of Response Times

Van der Linden (2007) mentioned an important notion from the reaction-time research, when working on a task, a subject has the choice between working faster with lower accuracy and working slower with higher accuracy. Thissen model and its variations (see Ferrando and Lorenzo-Seva (2007) and Ranger and Kuhn (2012)) typically represent such trade-off between speed and accuracy. In the same line we propose the response

time below in computerized testing situation,

$$\log(R_{i,t,s}) = \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s}. \tag{2.4}$$

Here, μ_i reflects the average response time for i-th respondent in general. $\nu_{i,t}$ implies the variation of the speed of the respondent i at the t-th day, with the negative sign indicating the slower the speed is, the more time needed to spend on the exam. We further assume the speed for an examinee will not change much during one day, thus the index of the speed only varies according to individuals and days. Let $\nu_{i,t}$ follow $\mathcal{N}(0, \kappa_i^{-1})$, with an individual -specific precision parameter κ_i for the variation and the mean centering at zero for ensuring identifiability in presence of μ_i. In the third term, $\theta_{i,t} - a_{i,t,s}$ indicates the distance between the ith person's ability on the tth day and the difficulty level for the s-th test on that day; $L(x)$ is a function to characterize the relationship between the distance and the response time; and β is a regression coefficient to adjust the influence of the distance function to the response time. Based on the intuition, the mechanism that controls the influence of the distance function to the response time is more or less the same among different tests and individuals, thus β is assumed to be a common parameter across different individuals and tests. $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$ is a residual term with a common precision parameter of ϱ to borrow strength of the data across different tests and individuals. Although ϱ varying across may be an alternative, such an assumption might cause identifiability issue with precision parameter κ_i when we encounter the

situation that an examinee only takes one test per day. For $L(x)$, we have two popular choices from the current literature. They are, $L(x) = x$, the monotone relationship and the other is $L(x) = |x|$, the inverted-U shaped relationship. In our case, we believe the theory of I-U shaped linkage (as suggested by Wang and Zhang (2006)) is applicable especially in the context of multi-time test takers. We also justify the relationship based on empirical findings from MetaMetrics data. Later in Chapter 2 we address model selection issues and we compare the fit of I-U shaped linkage with its competing linkage, namely monotone shaped. From the fit comparison results as demonstrated in Chapter 2, we find I-U shaped linkage choice is justified.

2.2.2 Second Stage: System Equations in the DIR-RT Models

Following the idea of Wang et al. (2013), we combine both parametric growth models and Markov chain models for modeling an individual's ability growth over time. Then, the model to describe his/her current ability $\theta_{i,t}$ is,

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}. \tag{2.5}$$

The first term in (2.5) denotes the ability at the previous time point, $\theta_{i,t-1}$. The second term represents a parametric growth model with c_i as the average growth rate of the i-th person's ability over time, where $\Delta^+_{i,t} = \min\{\Delta_{i,t}, T_{\max}\}$ is the time lapse between two test dates for ith invididual (i.e., $\Delta_{i,t}$) but truncated by a pre-specified maximum time

interval $T_{\max}$ ($T_{\max} = 14$ used in the application); ρ is the parameter to control the rate of one's growth, which reduces the growth rate when the ability becomes mature. Note ρ is known from empirical experiments in MetaMetrics datasets (details on identifiability issues between ρ and c_i, if ρ is treated unknown , are discussed in Wang et al. (2013)). Last $w_{i,t}, \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$ represents the random component of the change in the i-th person's ability on the t-th day with ϕ being a common unknown parameter to borrow information and to avoid a substantial risk of confounding in the likelihood between δ_i's and $\phi^{-1}\Delta_{i,t}$ when the time lapse between tests for the student are equally spaced. This assumption of $w_{i,t}$ presumes that the changes of one's ability change is much more uncertain, if he/she is absent for a long period. The system equation (2.5) can also be rewritten as a first-order Markov process (see *Step 2* of Appendix A), which is beneficial for conducting MCMC later.

2.2.3 A Summary of DIR-RT Models

To summarize, the proposed one-parameter DIR-RT models have two-stages, the 1st stage is composed of observation equations, while the 2nd stage is composed of system

equations,

$$
\begin{aligned}
\textit{2nd stage:} \qquad & \theta_{i,t} = \theta_{i,t-1} + c_i(1-\rho\theta_{i,t-1})\Delta^{+}_{i,t} + w_{i,t}, \\
\textit{1st stage:} \qquad & \Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}, \epsilon_{i,t,s,l}) \\
& \qquad = \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}, \\
& \log(R_{i,t,s}) = \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s},
\end{aligned}
$$

where $R_{i,t,s}$ and $X_{i,t,s,l}$ are observed; $a_{i,t,s}$'s, ρ, $\Delta^{+}_{i,t}$'s and $\Delta_{i,t}$'s are known and $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with known σ^2. Moreover, we have the following distribution assumptions. $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$ and $\nu_{i,t} \sim \mathcal{N}(0, \kappa_i^{-1})$. Here, $L(\theta_{i,t} - a_{i,t,s})$ either equals to monotone relationship, i.e., $(\theta_{i,t} - a_{i,t,s})$ or inverted U shape, i.e., $|\theta_{i,t} - a_{i,t,s}|$.

2.3 Statistical Inference and Bayesian methodology

As discussed in the section 1.5 we note that a frequentist's approach implementing Expectation Maximization (EM) or some version of that or marginalized maximized likelihood estimators (MML) is almost next to impossible to compute due to extremely complex nature of likelihood. In addition, some parameter spaces are restricted, which renders the EM or MML method extremely intractable. On top of that standard error estimates of the estimators are not very reliable. On the other hand Bayesian methodology

not only simplifies modeling and estimating the uncertainties, thanks to advancements in MCMC techniques, Bayesian computation is way simpler and easy to be extendable to other complex variants of the model. Next we describe and implement a fully (in contrast to those empirical Bayes implementation or similar implementation) Bayesian methodology.

2.3.1 Prior Distribution for the Unknown Parameters

Prior choice is crucial in any Bayesian analysis. In absence of expert's knowledge or historical information, objective priors are used for the unknown parameters to avoid the large impacts of priors on the inference and to have some good frequentist properties (Berger (2006)). Whenever, there are scientific knowledge available, we instead incorporate such information into the prior specification.

Following these rules, a natural choice for the prior of one's initial latent ability is $\theta_{i,0} \sim \mathcal{N}(\mu_{G_{j_i}}, V_{G_{j_i}})$, where $\mu_{G_{j_i}}$ and $V_{G_{j_i}}$ are the mean and the variance of the subpopulation (j) to which an individual i belongs. Since c_i's in the system equation (2.5) is the average growth rate and usually a learning rate in educational context is positive, then we choose the prior for c_i as

$$\pi(c_i) \propto \mathbf{1}(c_i \geq 0), \text{ for all } i,$$

where $\mathbf{1}(\cdot)$ is an indicator function. For the precision parameter of the speed (κ_i) and the

random error (ϱ), we utilize the usual scale objective prior, i.e., $\pi(\varrho) \propto 1/\varrho,$ and $\pi(\kappa_i) \propto 1/\kappa_i$ for all i. However, for the prior choice of scale parameters δ_i's, τ_i's and ϕ, we make the same choices as that of Wang et al. (2013) because of the requirement of posterior propriety, which assign them objective priors $\pi(\phi) \propto 1/\phi^{3/2},\ \pi(\delta_i) \propto 1/\delta_i^{3/2},$ and $\pi(\tau_i) \propto 1/\tau_i^{3/2}$ for all i. Further, a natural choice of the objective prior for μ_i, the average response time of each individual, is a constant prior, $\pi(\mu_i) \propto 1,$ for all i. Similarly, assume the prior of β is $\pi(\beta) \propto 1$. Although intuitively, the regression coefficient β in the observation equation of response times with a negative sign makes more sense, we let the real data help us determine the value and the sign of β.

2.3.2 Posterior Distribution and Data Augmentation Scheme

Using the fact that a standard logistic distribution can be expressed as a scale mixture of normals (Andrews and Mallows (1974)), one can write the density of Y, assuming Y follows a logistic distribution with location parameter 0 and scale $\pi^2/3$, as follows,

$$f(y) = \frac{e^{-y}}{(1+e^{-y})^2} = \int_0^\infty \left[\frac{1}{\sqrt{2\pi}}\frac{1}{2\nu}\exp\left\{-\frac{1}{2}(\frac{y}{2\nu})^2\right\}\right]\pi(\nu)d\nu\,, \tag{2.6}$$

where ν has the Kolmogorov-Smirnov(K-S) density,

$$\pi(\nu) = 8\sum_{\alpha=1}^{\infty}(-1)^{(\alpha+1)}\alpha^2\nu\exp\{-2\alpha^2\nu^2\}, \nu \geq 0.$$

Note that the density in square brackets in (2.6) is $\mathcal{N}(0, 4\nu^2)$.

Applying the data augmentation idea (Tanner and Wong (1987)), a latent variable $Y_{i,t,s,l}$ can be introduced for each response variable $X_{i,t,s,l}$, where $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}, 4\nu^2_{i,t,s,l})$ and $\Pr(X_{i,t,s,l} = 1|\theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) = \mathrm{P}(Y_{i,t,s,l} > 0|\theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l})$. Let us define $X_{i,t,s,l} = 1$ if $Y_{i,t,s,l} > 0$ and $X_{i,t,s,l} = 0$ otherwise, and the introduction of $Y_{i,t,s,l}$ can facilitate the MCMC computation although it introduces more unknowns. Since $\epsilon_{i,t,s,l} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, marginalizing out it results in $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, 4\nu^2_{i,t,s,l} + \sigma^2)$. Then, the one-parameter DIR-RT models (2.3), (2.4) and (2.5) can be rewritten as

$$\begin{aligned}
\theta_{i,t} &= \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}, \\
\log(R_{i,t,s}) &= \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s}, \\
Y_{i,t,s,l} &= \theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \xi_{i,t,s,l},
\end{aligned}$$

where $\xi_{i,t,s,l} \sim \mathcal{N}(0, \psi^{-1}_{i,t,s,l})$ with $\psi^{-1}_{i,t,s,l} = 4\gamma^2_{i,t,s,l} + \sigma^2$ and $\gamma_{i,t,s,l} \sim$ K-S distribution, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, $\nu_{i,t} \sim \mathcal{N}(0, \kappa_i^{-1})$ and $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$.

Define $\theta = (\theta_1, \cdots, \theta_n)'$ with $\theta_i = (\theta_{i,0}, \theta_{i,1}, \cdots, \theta_{i,T_i})'$; $c = (c_1, \cdots, c_n)', \tau = (\tau_1, \cdots, \tau_n)', \delta = (\delta_1, \cdots, \delta_n)', \mu = (\mu_1, \cdots, \mu_n)'$ and $\kappa = (\kappa_1, \cdots, \kappa_n)'$; $Y = \{Y_{i,t,s,l}\}, \gamma = \{\gamma_{i,t,s,l}\}$ and $X = \{X_{i,t,s,l}\}$; $\varphi = \{\varphi_{i,t}\}$ and $\nu = \{\nu_{i,t}\}$; $\log R = \{\log R_{i,t,s}\}, \eta = \{\eta_{i,t,s}\}$ and $\eta^*_{i,t} = (\eta_{i,t,1}, \cdots, \eta_{i,t,S_{i,t}-1})'$; where $l = 1, \cdots, K_{i,t,s}, s = 1, \cdots, S_{i,t}, t = 1, \cdots, T_i$ and $i =$

$1, \cdots, n$. Given the data $(X, \log R)$, the joint posterior density of $(\theta, Y, c, \tau, \varphi, \eta, \phi, \beta, \nu, \mu, \varrho, \gamma)$ of our proposed DIR-RT models is

$$
\begin{aligned}
& \pi(\theta, Y, c, \tau, \varphi, \eta, \phi, \beta, \nu, \mu, \varrho, \gamma \mid X, \log R) \\
\propto\ & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} \pi(\gamma_{i,t,s,l}) \right\} \left\{ \prod_{i=1}^{n} \pi(\theta_{i,0}) \pi(c_i) \pi(\delta_i) \pi(\tau_i) \pi(\kappa_i) \right\} \pi(\beta) \pi(\phi) \pi(\varrho) \\
\times\ & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} [\mathbf{1}(Y_{i,t,s,l} > 0) \mathbf{1}(X_{i,t,s,l} = 1) \mathbf{1}(Y_{i,t,s,l} \leq 0) \mathbf{1}(X_{i,t,s,l} = 0)] \right. \\
\times\ & \left. \sqrt{\frac{\psi_{i,t,s,l}}{2\pi}} \exp\left(-\frac{\psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2} \right) \mathbf{1}\left(\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s} \right) \right\} \\
\times\ & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \sqrt{\frac{\delta_i}{2\pi}} \exp\left(-\frac{\delta_i \varphi_{i,t}^2}{2} \right) \right\} \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \left(\frac{\tau_i}{2\pi} \right)^{(S_{i,t}-1)/2} \exp\left(-\frac{\tau_i \eta_{i,t}^{*\prime} \Sigma_{i,t}^{-1} \eta_{i,t}^{*}}{2} \right) \right\} \\
\times\ & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \sqrt{\frac{\phi}{2\pi\Delta_{i,t}}} \exp\left(-\frac{\phi[\theta_{i,t} - \theta_{i,t-1} - c_i(1 - \rho\theta_{i,t-1})\Delta_{i,t}^{+}]^2}{2\Delta_{i,t}} \right) \right\} \\
\times\ & \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{s=1}^{S_{i,t}} \sqrt{\frac{\varrho}{2\pi}} \exp\left(-\frac{\varrho(\log(R_{i,t,s}) - \mu_i + \nu_{i,t} - \beta L(\theta_{i,t} - a_{i,t,s}))^2}{2} \right) \\
\times\ & \prod_{i=1}^{n} \prod_{t=1}^{T_i} \sqrt{\frac{\kappa_i}{2\pi}} \exp\left(-\frac{\kappa_i \nu_{i,t}^2}{2} \right), \qquad (2.7)
\end{aligned}
$$

where $\Sigma_{i,t}^{-1} = \mathbf{J}_{S_{i,t}-1} + \mathbf{I}_{S_{i,t}-1}$, with $\mathbf{J}_{S_{i,t}-1}$ being a $(S_{i,t}-1) \times (S_{i,t}-1)$ unit matrix, $\pi(\theta_{i,0})$, $\pi(c_i)$, $\pi(\delta_i)$, $\pi(\tau_i)$, $\pi(\kappa_i)$, $\pi(\beta)$, $\pi(\phi)$, and $\pi(\varrho)$ are the priors specified in 3.1; and $\pi(\gamma_{i,t,s,l})$ is the K-S density. The proof of posterior propriety of DIR-RT models closely follows from a simple extension of Appendix C in Wang et al. (2013) for DIR models. The major difference of DIR-RT and DIR models is the part that conjointly models response times with item responses. Since the logarithm of response times is modeled

as a normal regression model, the well established facts for the posterior propriety of normal regression models in Bayesian literature for the objective priors specified in 3.1 can be coupled with results from Appendix C in Wang et al. (2013) to show the posterior propriety of DIR-RT models.

2.3.3 MCMC Computation of DIR-RT Models

The computation is carried out by MCMC scheme that samples from the posterior (2.7) via block Gibbs sampling schemes. The difficulty of the sampling scheme is to draw the posterior distribution of latent ability $\theta_i = (\theta_{i,0}, \cdots, \theta_{i,T_i})'$ for each individual i, for $i = 1, \cdots, n$, where coordinates of θ_i are typically high dimensional and strongly correlated. When $L(x) = x$, using the novel data augmentation idea, the proposed model is transformed so that θ_i could be block sampled – within a Gibbs sampling step conditional on the other parameters – by the highly efficient forward filtering and backward sampling algorithm (West and Harrison (1997)). However, if $L(x) = |x|$, the computation becomes more challenging as θ_i cannot be drawn as a block. Instead, we utilize the fact that given all the other unknowns, the full conditional distribution of each coordinate $\theta_{i,t}$ follows a mixture of truncated Gaussians, so that $\theta_{i,t}$ can be drawn one at a time, thus being integrated as an extension of MCMC chain.

The details of MCMC steps are given in Appendix A. The Gibbs sampling starts at *Step 1* in Appendix A, with initial values for $\theta^{(0)}$, $c^{(0)}$, $\phi^{(0)}$, $\varphi^{(0)}$, $\eta^{(0)}$, $\delta^{(0)}$, $\tau^{(0)}$, $\gamma^{(0)}$, $\mu^{(0)}$, $\nu^{(0)}$ and $\beta^{(0)}$, then loops through *Step 15* in Appendix A, until the MCMC has

converged. The initial values chosen in the applications were $\theta^{(0)} = \vec{0}$, $c^{(0)} = \vec{0}$, $\phi^{(0)} = 1$, $\varphi^{(0)} = \vec{0}$, $\eta^{(0)} = \vec{0}$, $\delta^{(0)} = \vec{1}$, $\tau^{(0)} = \vec{1}$, $\gamma^{(0)} = \vec{1}$, $\mu^{(0)} = \vec{1}$, $\nu^{(0)} = \vec{0}$ and $\beta^{(0)} = 0$. The convergence was evaluated informally by looking at trace plots.

Then, statistical inferences are made straightforward from the MCMC samples. For example, an estimate and 95% credible interval (CI) for the latent trajectory of one's ability $\theta_{i,t}$ can be plotted from the median, 2.5%, and 97.5% empirical quantiles of the corresponding MCMC realizations. In examples, ability will be graphed as a function of t, so that the dynamic changes of an examinee is apparent.

2.4 Simulation Study

To validate the inference procedure and compare the benefits by jointly modeling response times with item responses, a simulation study was conducted with similar set-up as laid out in section 4 of Wang et al. (2013). To save the space, we only illustrate the situation when the simulation model for response times indeed follows an inverted U-shaped linkage for the distance of ability-difficulty with response times. Similar results can be obtained if we had proceeded with monotone shape linkage. The simulation method considers multiple individuals taking a series of tests scheduled at individually-varying and irregularly-spaced time points.

2.4.1 DIR-RT Models Simulation

Following the simulation study of DIR models in Wang et al. (2013), assume there are 10 individuals, each of them has taken four tests on 50 different test dates, where each test contains 10 items. The specification means $K_{i,t,s} = 10$, for $s = 1, \cdots, S_{i,t}$, $t = 1, \cdots, T_i$, $i = 1, \cdots, n$ with $S_{i,t} = 4$, $T_i = 50$ and $n = 10$. Let time lapse between two consecutive test dates be Δ_{it}, where $\Delta_{it} = t + 10$ if $t \leq T_i/2$ or $\Delta_{i,t} = t - 10$ otherwise, creating a irregularly spaced gap between two test dates.

In order to do the comparison of DIR-RT models with DIR models, we assign same values of the parameters, ϕ, δ_i, τ_i, c_i used in Wang et al. (2013), where $\phi = 1/0.0218^2$, leading standard deviation of $w_{i,t}$ in system equation (2.5) is $0.0218\sqrt{\Delta_{i,t}}$ and the values of δ_i, τ_i, c_i are specified in Table 2.1. For the modeling part of response times, the parameter values of κ_i and μ_i are shown in Table 2.1,

i	1	2	3	4	5	6	7	8	9	10
c	0.0055	0.0065	0.0026	0.0037	0.0061	0.0047	0.0035	0.0043	0.0039	0.0015
δ	2.0408	1.3333	1.8182	1.2346	1.5873	1	2.2222	1.0526	1.1494	2
τ	4	3.1250	4.3478	2.7027	3.7037	2.8571	4	2.2222	9.0909	4.5455
κ	2.3256	1.5873	1.6949	0.5495	1.2658	0.9346	1.3889	1.8182	2.7027	1.2195
μ	1.6	1.47	1	1.92	1.45	1.73	1.5	1.35	0.81	1.23

Table 2.1: Values of the parameters used in DIR-RT simulation

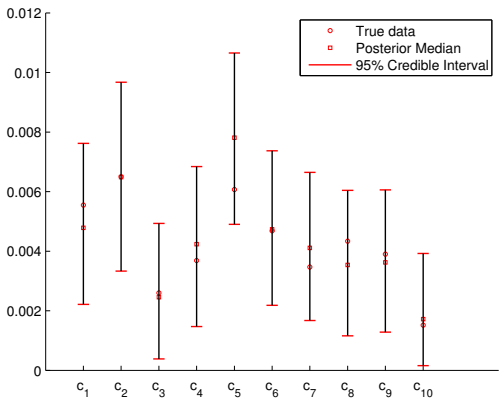
$\beta = -0.17$ and $\varrho = 1.25$. The parameters of DIR-RT models are chosen in such a way that they are in order of same magnitude to mimic the real data from MetaMetrics company.

Consider the inverted U shape linkage for $L(x)$, i.e., $L(x) = |x|$. Simulation proceeds

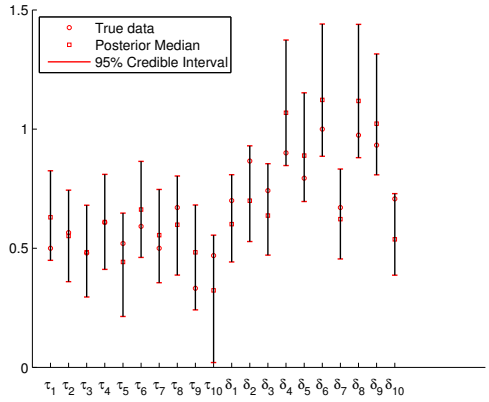
by simulating random effects or latent variables using the assigned parameter values above for DIR-RT models. Once we get the simulated values for $\theta_{i,t}$ using *2nd stage* model in 2.3, the test difficulties, $a_{i,t,s}$ in *1st stage* model is set to be $\theta_{i,t} + \zeta^*$, where ζ^* is a random variable with uniform distribution on $(-0.1, 0.1)$. The values of $\epsilon_{i,t,s,l}$ are drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.7333$ and we choose $\rho = 0.1180$. Notice the values of σ and ρ are used in MetaMetric application. The dichotomous data of item responses and continuous data of response times generated from the simulation are our observations, and the Bayesian methodology from section 3 is implemented in estimating the model parameters of DIR-RT models.

The parameters are estimated through posterior median calculated from their corresponding MCMC samples. Each MCMC was run for 50,000 iterations with a 25,000 burn-in period. Figure 2.1 (a)-(d) give posterior median estimates (red squares) along with 95% CIs (red bars) of c, $\tau^{-1/2}$, $\delta^{-1/2}$, $\kappa^{-1/2}$ and μ, respectively and illustrate their true values (black dot). Clearly from Figure 2.1, the true values of those parameters are contained within their corresponding 95% CIs. For the posterior median estimates of parameters $\phi^{-1/2}$, $\varrho^{-1/2}$, β are 0.0190, 0.9075, -0.1815, respectively, with their corresponding 95% CIs being $[0.0159, 0.0229]$, $[0.8753, 0.9427]$, and $[-0.7028, -0.0071]$, all of which contain their true values.

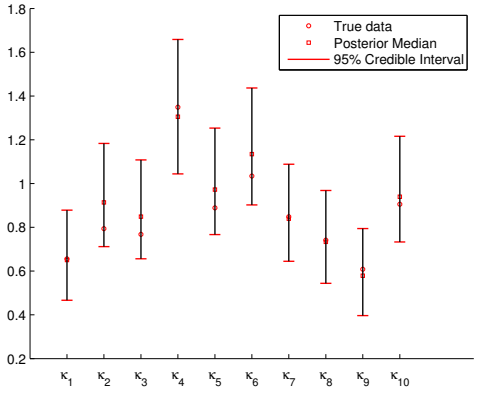
Next, we turn our focus to the primary interest for estimating latent ability trajectories. Figure 2.2 (a)-(d) illustrate four types of growth curves in our simulation, where (a) θ_1 represents an individual with steady growth; (b) θ_2 indicates an individual with



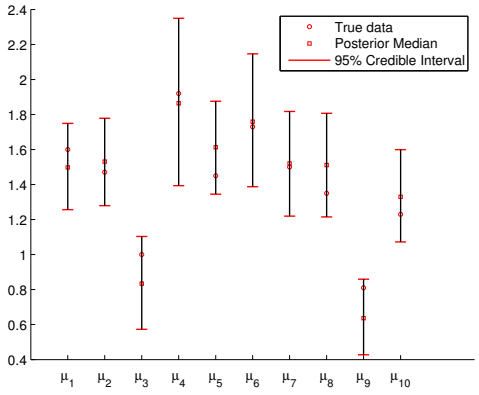
(a) c_i's



(b) $\tau_i^{-1/2}$'s and $\delta_i^{-1/2}$'s



(c) $\kappa_i^{-1/2}$'s



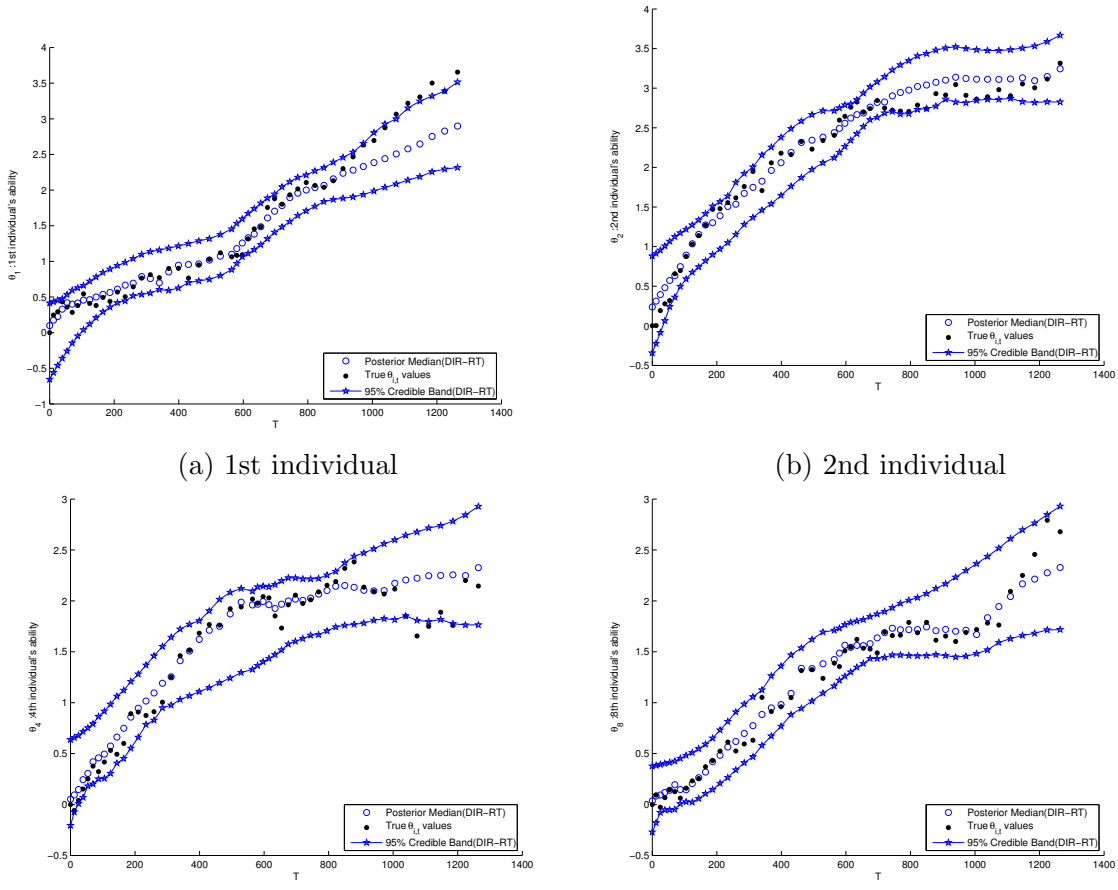
(d) μ_i's

Figure 2.1: Posterior Summary of c_i's, $\tau_i^{-1/2}$, $\delta_i^{-1/2}$'s, $\kappa_i^{-1/2}$'s, and μ_i's, where red circles represent true values, red squares are the posterior median estimates and red bars indicate 95% CIs.

increasing growth but nearly flat region at the end; (c) θ_4 shows an individuals with interrupted growth (with true ability drops in certain period); (d) θ_8 displays monotonic growth with decreasing growth rate in the middle. In Figure 2.2, the true ability curves (black dots) have been plotted along with our posterior median estimates of ability (blue circles) and their corresponding 95% credible band (starred lines). Notice in each subfigure, very small proportion of true values are outside of 95% credible bands.

To better assess how well the Bayesian methodology actually captures the truth, we use the coverage probability (CP), represented by the frequency of true values falling in the corresponding CIs over different MCMC runs. To evaluate CPs, we conduct the simulation with the same setting specified eariler but with 10 different sets of seeds for random number generations. The CPs for ϱ, β, ϕ are 90%, 100% and 90%, respectively and the average CPs over all individuals for κ_i's, c_i's, τ_i's, δ_i's and μ_i's are 94%, 96% , 92%, 95% and 95%, respectively. In addition, the average CPs for one's ability across different time, i.e., for $\theta_1, \cdots, \theta_{10}$, are 97%, 95.6%, 96.2%, 97.4%, 94.6%, 98.4%, 97.8%, 95.2%, 97.8% and 96.0%. Thus, while the inferential method is Bayesian, it seems to yield sets that have good frequentist coverage.

Figure 2.3 displays the growth curve of two selected individuals (i.e., θ_2 and θ_6), where the statistical inference is based on the simulated example in 2.4. For other individuals, results are the similar and to save the space, we omit the plots for others. In Figure 2.3 (a) and (b), 95% CIs of DIR models (dotted red lines) encompass 95% CIs of DIR-RT



(a) 1st individual

(b) 2nd individual

(c) 4th individual

(d) 8th individual

Figure 2.2: The latent trajectory of one's ability growth, where black dots, blue circles and starred lines represent true ability, the posterior median estimates and the 95% credible bands, respectively.

models (starred blue lines), both 95% CIs contain the true values (black dots). The average length of 95% credible band of ability estimates for DIR-RT models is much shorter than that of DIR models (0.6454 vs 1.0370 for θ_2 and 0.6772 vs 1.1401 for θ_6, respectively). In addition, notice that in Figure 2.3, both graphs of DIR-RT for ability estimates (blue circles) adhere more closely to true ability (black dots) in relative to DIR ability estimates (red dots). The average mean squared distance between the truth and the posterior median ability estimates over time for DIR-RT models are 0.0240 for θ_2 and 0.0187 for θ_6, in comparison to that of DIR models are 0.0711 for θ_2, and 0.0653 for θ_6. The results illustrate that by incorporating response times, we can largely improve the precision and remarkably reduce the bias of the estimates of one's ability trajectory.

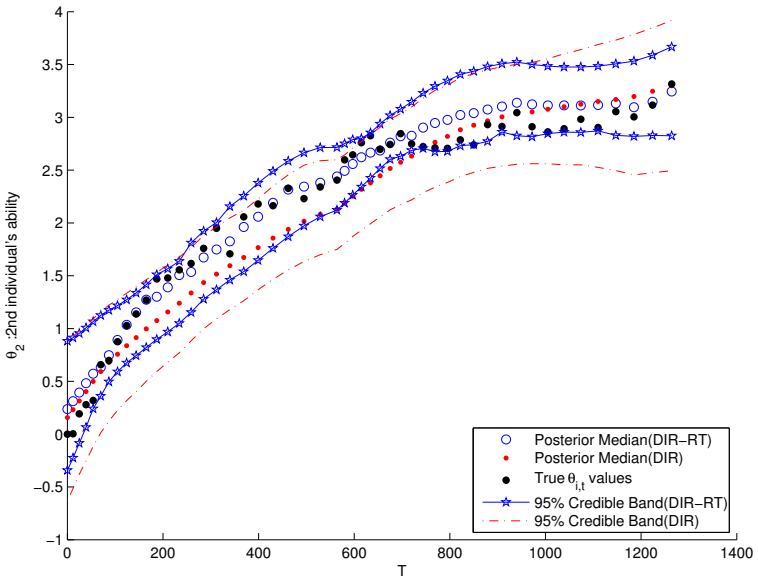
2.5 MetaMetric Testbed Application

For illustration purpose, we randomly select a sample of 25 individuals from MetaMetrics testbed datasets. There are different characteristics for each student as shown in Table 2.2.

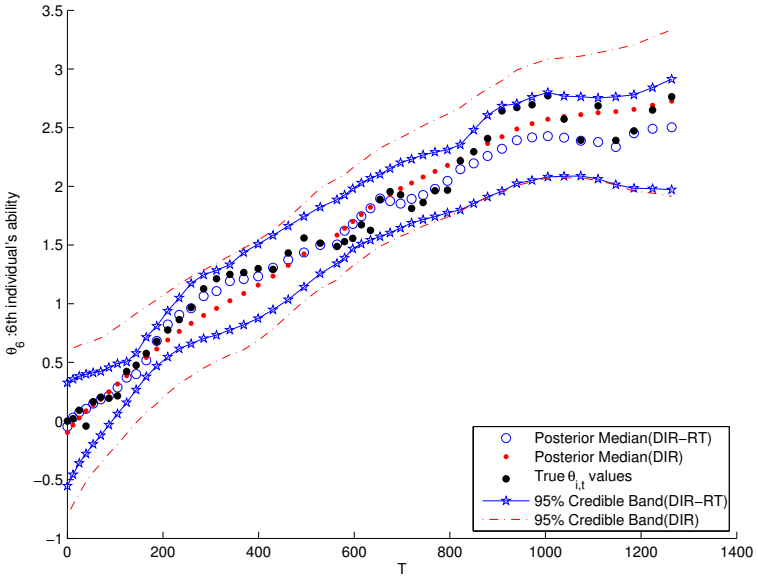
	Total Tests	Days	Max. Tests/Days	Range of Items/Test	Max. Gap	Initial Grade
No.1	150	74	9	4-22	79	4
No.2	203	128	15	6-24	107	2
No.3	211	107	9	5-24	79	3

Table 2.2: Characteristics of the first 3 individuals randomly sampled from the MetaMetrics data

To save space, we only show the details for the first three individual we selected. The



(a) 2nd individual



(b) 6th individual

Figure 2.3: The comparison of ability estimates between DIR-RT and DIR models, where black dots, blue circles, red dots represent true mean ability, DIR-RT ability estimates, DIR ability estimates respectively; starred-lines (blue) and dash (red) lines represent 95% credible bands for DIR-RT and for DIR respectively.

primary focus of this application is to study the following goals: 1) assessing the appropriateness of the local independence assumption for this type of data; 2) understanding the growth in ability of students, by retrospectively producing the estimated growth trajectories of their abilities in the study; 3) investigating whether the proposed linkage, that is inverted-U shaped linkage between response times and the distance of ability-difficulty to model students' behaviors and psychology in the exam can be justified in light of the data.

2.5.1 Using Lindley's Method to Test the Significance of I-U Shaped Linkage

The regression slope β plays a key role in controlling the influence of the ability-difficulty distance function to the response time. It is easy to note that DIR models are nested with DIR-RT models. When β becomes 0 collateral information due to response time model does not add to improve the estimates of response model. Thus significance of I-U shaped linkage based on response time model depends on the value of regression coefficient β. Thus, we are interested in testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, since $\beta = 0$ implies the distance between ability and difficulty does not affect the time that individual spends on a test and the corresponding linkage function $L(\cdot)$ can be ignored. Lindley's method (Lindley (1965), Section 5.6), advocated by authors such as Zellner (1971), is an ad hoc way to test this. According to Lindley's method, one rejects the

hypothesis : $\beta = 0$ at the α level of significance if the $100(1-\alpha)\%$ highest posterior density interval does not include 0. The posterior density of β is in bell shapes (Please see the left histogram in Figure 3.4). So $100(1-\alpha)\%$ highest posterior density interval is the same as $100(1-\alpha)\%$ CI.

Model(β)	Inverted U-shape
PM	-0.2305
95% CI	$(-0.2940, -0.1571)$
99% CI	$(-0.3105, -0.1345)$

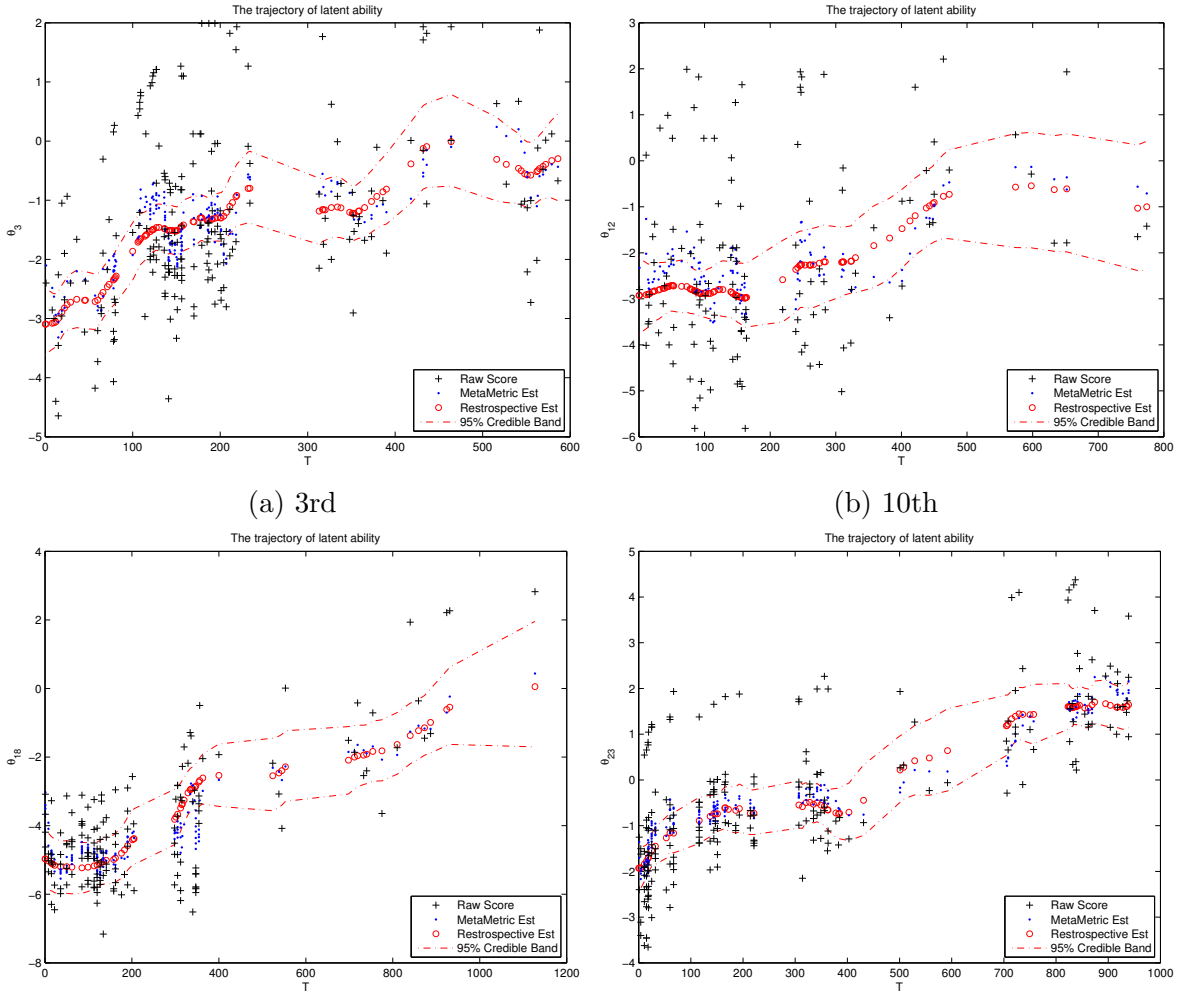
Table 2.3: The posterior summary of β under inverted U-shape, where 'PM' in the table is the abbreviation for 'posterior median'.

It is evident from Table 2.3 that $\beta = 0$ is rejected at both $\alpha = 1\%$ and $\alpha = 5\%$ for inverted U-shape. This phenomenon strongly suggests that there is a inverse relationship (negativity of β value) between inverted U shaped linkage and response time. We shall re-visit the same model selection issue in Chapter 2 and I-U shaped linkage would be validated while compared with other competing alternatives with the help of a novel model selection criterion.

2.5.2 Retrospective Estimation of Ability Growth Under I-U Shaped Linkage

As Lindley's Method supported the choice of inverted U-shape linkage for the analysis of our proposed DIR-RT models for MetaMetrics data, we are going to use inverted U-shape through the rest of the paper accomplish other two goals mentioned at the beginning of this section. Figure 2.4 presents a retrospective analysis of the reading ability for 3rd,

10th, 18th and 24th individuals (using all data recorded of each individual during the study period). In Figure 2.4, the red circles are the posterior median estimates of each individual's ability, the red dash lines correspond to the 2.5% and 97.5% quantiles of the posterior distributions of the abilities, and the black plus points are raw scores. Similar as Wang et al. (2013), we find all these growth trajectories have an overall increasing trend but such kind of growth can be interrupted. In particular, when there is a large time gap between subsequent tests, the ability appears to drop for some individuals, which is clearly seen from Figure 2.4. Some natural explanations might be that during vacations, a student may not read and could actually lose ability or they become less used to computerized tests after a long break.



(a) 3rd

(b) 10th

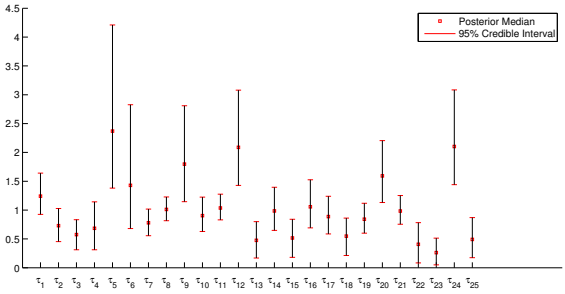
(c) 18th

(d) 24th

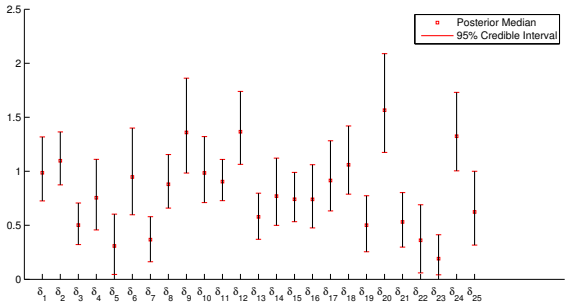
Figure 2.4: The posterior summary of the ability growth for θ_3, θ_{10}, θ_{18} and θ_{24}, where red circles, black plus and blue dots represent posterior median estimates of the ability, raw score and MetaMetric estimates, respectively and red dash lines represent 95% CIs.

Figure 2.5 and Figure 2.6 together give the posterior summaries (i.e., the posterior median (red squared rectangles dot) and 95% CI (red bars at two ends)) of the average growth rates c_i's, the standard deviations of test random effects $\tau_i^{-1/2}$'s, the standard deviations of the daily random effects $\delta_i^{-1/2}$'s, the standard deviations of speediness, $\kappa_i^{-1/2}$, and the average response time for each individual, μ_i, for $i = 1, \cdots, 25$. Moreover, the estimated posterior median of $\phi^{-1/2}$ is 0.0708 and its 95% CI is $[0.0608, 0.0831]$ and the result of β is shown in the table of Figure 2.3.

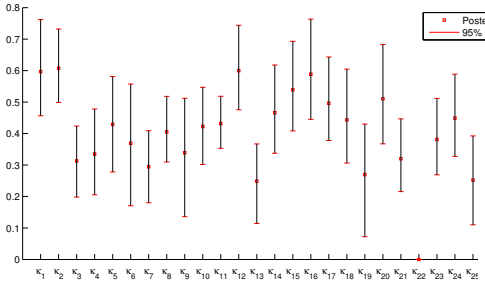
Figures 2.5 (a)-(b) show that the standard deviations of test and daily random effects are almost all quite large with 95% CIs and are well separated from zero. Recall that these were included in the model to account for a possible lack of the local independence; the evidence is thus strong that the local independence is, indeed, not tenable for this data and that both types of random effects are present for most individuals. Additionally, Figure 2.5 (c) illustrates that speediness of individuals is different on the daily basis except that of individual 22nd (almost steady during the studying period). Moreover, there are clearly some patterns of the variation of speediness for individuals, some of them, the difference of their speediness on a daily basis is more crucial than that of the others. The variation in the average response time in Figure 2.5 (d) suggests also some individuals take longer time to finish a test than others. As well, it is not surprising the average growth rates are quite different for each individual as shown in Figure 2.6.



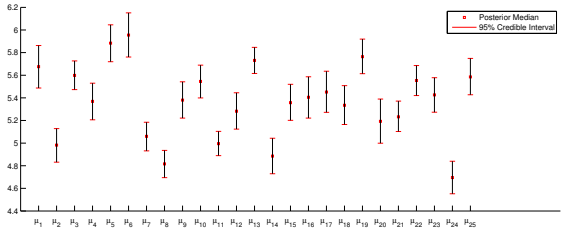
(a) The posterior median and 95% CI of δ



(b) The posterior medianand 95% CI of τ



(c) The posterior median and 95% CI of κ



(d) The posterior median and 95% CI of μ

Figure 2.5: Posterior summary of c, $\tau^{-1/2}$, $\delta^{-1/2}$, $\kappa^{-1/2}$ and μ

2.6 Discussion

From our simulation study, we noticed that incorporation of response time into the item response model for the analysis of individually varying and irregular spaced longitudinal observations has both significantly improved the precision and reduced the bias for the ability estimation for the proposed models. Using DIR-RT models to analyze MetaMetric datasets, the results further support findings of Wang et al. (2013). For example, the evidence of violation of the local dependence assumption is generally strong in DIR-RT models, and use of test and daily random effects to model the local dependence seems to be necessary and successful. The retrospective analysis of ability estimation is of

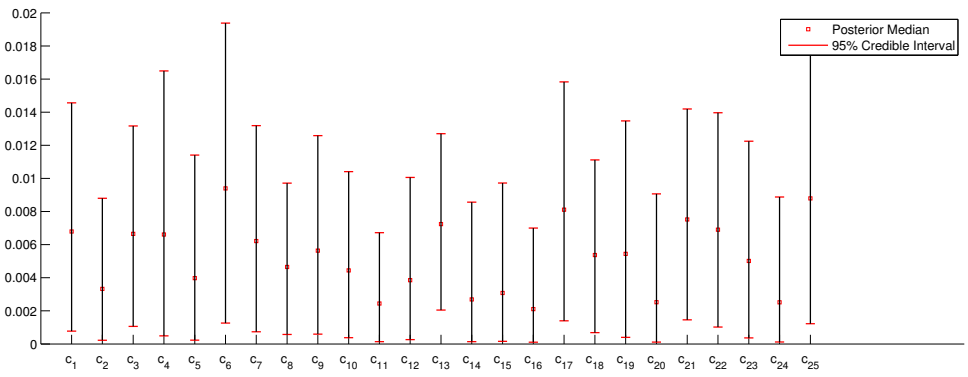


Figure 2.6: The posterior median and 95% CI of c

considerable use in understanding population behavior, such as the frequently observed drops in ability after a long pause in testing.

The result favors the inverted U-shape linkage, which is quite important and meaningful conclusion, since it supports that in such tests the psychology of students in a test is to spend more time on items or questions that match their ability levels and spend less time on those, either too easy or too hard. In the next Chapter we formally develop an approach to compare models within DIR-RT framework and show that IU shaped linkage is the best choice with respect to this newly developed measure, thus confirming the results we established in this Chapter. We would like to add here that using response time as an extra information helps improve the estimates of response model also.

Many extensions are possible, such as extensions to two-parameter and three-parameter DIR-RT models (as mentioned before also). Figure 2.5 clearly illustrates some patterns among individuals for the average growth rate c_i's, the variation of speediness κ_i's, and the average response time μ_i's. Next step could also be to consider , using either

model-based or distance-based clustering methods, to further investigate the presence of groups of different some psychological behavioral patterns. It is anticipated that the clustering can further help teachers better assist their students and achieve the goals of personalized education.

Chapter 3

Model Selection in DIR-RT Framework

3.1 Introduction and Motivation

Though recent literature showed evidence in favor of I-U shaped models (Wang and Zhang (2006)), one can raise a question if this is truly the only alternative. Based on recent research work, it can be argued that DIR-RT models are quite flexible framework that tries to marry the two, DIR models and response time models in a reasonable way. A possible alternative to DIR-RT with I-U shaped linkage would be DIR-RT with monotone linkage, which is a quite popular linear linkage (please see Van der Linden (2007)), that has been around for some time. In addition, within DIR-RT framework this alternative would be slightly faster to compute, thanks to Forward Filtering Backwards Sampling (FFBS, West and Harrison (1997)) scheme that works very well within Gaussian family of distributions as opposed to mixture of truncated Gaussians. So there is computational trade-off possible between the two. This motivates us to search for the best model among

these options within DIR-RT framework.

3.1.1 Bayes Factor and DIC as Selection Criteria

One of the popular and simple-to-compute measures is to use posterior odds or Bayes factor (Jeffreys (1998)). They are equivalent as long as one has an idea of prior odds as posterior odds is the product of Bayes factor and prior odds. Let M denote the model we believe is true and let Θ be the parameter vectors present, which may include latent variables if the model permits so, then posterior odds is the ratio of two posterior predictive densities where Bayes factor is the ratio of prior predictive densities. In this method one chooses a model that has maximum posterior probability. Mathematically prior predictive density given a model is the following integral.

$$p(Y \mid M) = \int_{\Theta} p(Y \mid \Theta, M) p(\Theta \mid M) \partial\Theta, \quad p(\Theta \mid M) \text{ prior due to model M.}$$

As the dimensionality of the integral increases the computation gets harder. In addition the parameters may be over restricted spaces. In DIR-RT case this amounts to integrating over all latent variables like ability as well as γ's which would not have a closed form both in I-U shaped and in monotone case. Mathematical approximations (based on Laplace method or quadratic approximations) of the integral will be equally complex. Numerical approximation will be out of question because of high dimensionality as Monte Carlo based estimates would not be reliable due to insufficient sample

size.

Another very popular model selection criterion, called deviance information criterion (DIC), is usually recommended in complex models as it tries to take care of complexity by penalizing on what is called effective sample size. Since its first introduction by the seminal paper of Spiegelhalter, Best, Carlin, and Van Der Linde (2002), there have been many variants of DIC's such as conditional or complete data DIC and etc, mostly depending on the need for quick computations. In general, DIC can be defined in the following way

$$\begin{aligned} \text{DIC} \quad &= \overline{D(\boldsymbol{\theta})} + 2p_D \,[\text{set h(y)} = 1\,] \\ \text{or} \quad &= -4\mathbb{E}_{\theta|\mathbf{y}}[\log f(\mathbf{y}|\boldsymbol{\theta})] + 2\log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}), \end{aligned} \tag{3.1}$$

where deviance, $D(\boldsymbol{\theta}) = -2\log f(\mathbf{y}|\boldsymbol{\theta}) + 2\log h(\mathbf{y})$ and $\overline{D(\boldsymbol{\theta})}$, posterior mean deviance, $= -2\mathbb{E}_{\theta}[\log f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}] + 2\log h(\mathbf{y})$ and p_D is called effective number of parameters and is given as follows

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\tilde{\boldsymbol{\theta}}),\ \tilde{\theta}: \text{ posterior mean or mode }.$$

Here, θ denotes the vector of parameters. So the computation of DIC reduces to computation of mean log likelihood (also called integrated likelihood as no latent variable and/ or no augmented data is involved) under posterior distribution and in applied problem $\tilde{\theta}$ is taken as MAP. This DIC is usually denoted by DIC_2. In presence of latent

variables, say $\mathbf{Z}$, $f(y|\theta) = \int f(y,z|\theta)dz$, $f(y|\theta)$ is often referred to as observed data likelihood or integrated likelihood where as $f(y,z|\theta)$ is referred to as complete-data likelihood . But it may not be possible to integrate Z out analytically. So another variant of DIC, known as DIC_5, replaces $\mathbb{E}_\theta[\log f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}]$ by $\mathbb{E}_{\theta,Z}[\log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})|\mathbf{y}]$, that is estimated by posterior samples of (θ, Z), and $\log f(\mathbf{y}|\tilde{\boldsymbol{\theta}})$ by $\log f(\mathbf{y},\hat{\mathbf{Z}}|\hat{\boldsymbol{\theta}})$, where $(\hat{Z},\hat{\boldsymbol{\theta}})$ is joint MAP. In some cases complete-data like may not be easy to calculate but conditional likelihood could be easy to calculate. This leads to DIC_7, that replaces $\mathbb{E}_{\theta,Z}[\log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})|\mathbf{y}]$ by $\mathbb{E}_{\theta,Z}[\log f(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})|\mathbf{y}]$ and $\log f(\mathbf{y},\hat{\mathbf{Z}}|\hat{\boldsymbol{\theta}})$ by $\log f(\mathbf{y}|\hat{\mathbf{Z}},\hat{\boldsymbol{\theta}})$. Nevertheless, some recent studies have cautioned against the use of the DIC for comparing latent variable models. For instance, a popular model for count data with over dispersion is the Poisson-log normal model. Millar (2009) shows that the DIC obtained using conditional Poisson likelihood is inappropriate. Instead, the DIC calculated using the integrated likelihood seems to perform well. In this section we shall focus on integrated DIC or observed DIC as defined in 3.1. For "Monotone" model, integrated likelihood can be shown to have the following form after few algebraic steps with appropriate choices of α^*, C and Σ, which are only

functions of fixed parameters.

$$
\begin{aligned}
p(H|\Sigma,\gamma,S) &\sim N(\alpha^*, C^{-1}\Sigma) \\
p(Y|\Sigma,\gamma,S,LogR) &\sim N(\alpha^*_Y + (C^{-1}\Sigma)_{Y.R}(C^{-1}\Sigma)^{-1}_{R.R} \\
&\quad (LogR - \alpha^*_R), (C^{-1}\Sigma)_{YY.R}) \\
p(LogR|\Sigma,\gamma,S) &\sim N(\alpha^*_R, (C^{-1}\Sigma)_{RR}) \\
L(\boldsymbol{\theta}|X,LogR) &= \int_\gamma p(LogR|\Sigma,\gamma,S) \\
&\quad [\int_A p(Y|\Sigma,\gamma,S)dY]\, p(\gamma)d\gamma,\ \ A : I(x=1)(-\infty,0) + I(x=0)(0,\infty)
\end{aligned}
$$

It is evident that even for computationally the simpler case ("Monotone model" because of its conditional Gaussian structure) computation is almost next to impossible. Other numerical or analytical approximations would not work for the same reasons as mentioned in the earlier section.

3.1.2 Other Approaches

As an alternative to DIC, Bayesian χ^2 (Johnson (2004)) approach may be applicable. Unfortunately this entails computation of CDFs and inverse CDFs conditionally on parameters. Thus it does not seem to be easy to compute because of the presence of random effects, which need to be integrated out. We suggest an approach, motivated by the work of Yao, Kim, Chen, Ibrahim, Shah, and Lin (2015), which is to compute conditional DIC based on response time model only assuming the ability parameters are

estimated well.

3.1.3 Preview

In section 2 we introduce the new criterion based on conditional DIC methods and provide theoretical justifications behind that. In section 3 we provide some simulation study for the goodness of this criterion and we discuss its performance. In section 4, we apply this method to choose best models between two linkages in the context of MetaMetric testbed data and we re-visit choosing the right linkage for response models. Finally in section 5, we summarize the findings and discuss the limitations of the current study as well as some future works.

3.2 Partial DIC

We begin with the definition of DIC_7 (as in 3.1.1), which replaces the usual likelihood by conditional likelihood and approximates posterior deviance mean by averaging over posterior samples of joint of parameters and latent variables.

$$DIC_p = -4\mathbb{E}_{\Theta|\mathbf{y}}[\log f_P(\mathbf{logR}|\mathbf{\Theta})] + 2\log f_P(\mathbf{LogR}|\tilde{\mathbf{\Theta}}). \tag{3.2}$$

Here $\Theta = (\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \varrho)$ and $\tilde{\Theta}$ is joint MAP and f_P represents the partial density due to response time only, integrated over $\boldsymbol{\nu}$, conditioned on $\boldsymbol{\theta}$ under the model. Exact computation of the analytical expression is given in appendix B. The above definition

is motivated by the idea that conditionally treating ability is known DIR model is independent to response time(RT) model. So conditionally on ability the information due to change in response time models is only affected by the conditional likelihood due to response time as information on DIR model remains the same under both the models. So if ability is estimated reasonably well departures from the true models can be captured from conditional response time likelihood only.

3.3 Goodness of DIC_p as A Decision Rule: Simulation Study

To test how our proposed criterion performs in simulated data, we resort back to similar set-up as described in 2.4. We generate response and response-time data using the same parameter values as suggested in 2.4 and a specific true linkage. So we vary the seeds and the linkages to generate various data versions. It is worth noting here that value of regression coefficient, β, plays a critical role as it distinguishes these two models with two different linkages. If the value is close to 0 then response time model may be insignificant and both linkages may not be distinguishable from one other. Similarly higher values of β help distinguishing the two linkages. Therefore, during simulation study we need to be careful about the choices of β values. In this simulation study we worked with β = -1,-2 and -10. For each choice of β and $L(x)$ we worked with 10 runs of the data generating process. To study the goodness of the criterion, DIR-RT models with both linkages are

fitted to each of these data sets and partial DIC (DIC_p) is computed based on posterior samples for each of the linkage models based on the same data-set. A linkage model is selected based on smaller DIC_p value. Eventually we study misclassification rates to evaluate DIC_p.

3.3.1 Fitting DIR-RT Models on Simulated Data

For I-U shaped linkage it is demonstrated in section 2.4 . Modeling specifications can be summarized following 2.2.3.

$$
\begin{aligned}
\textit{2nd stage:} \qquad & \theta_{i,t} = \theta_{i,t-1} + c_i(1-\rho\theta_{i,t-1})\Delta^{+}_{i,t} + w_{i,t}, \\
\textit{1st stage:} \qquad & \Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}, \epsilon_{i,t,s,l}) \\
& \qquad = \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}, \\
& \log(R_{i,t,s}) = \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s}.
\end{aligned}
$$

Note that all the interpretations regarding parameters remain the same as described in 2.2.3. The only change is in the interpretation and in the definition of $L(x)$ function. In this case $L(x) = x$ as opposed to the I-U shaped linkage (where it was $L(x) = |x|$). This observation gives us a way to modify the full conditionals, given for IU linkage, in this case. First of all, full conditionals of the parameters and latent variables of the response model other than θ remain unchanged because of conditional independence as stressed in 3.2. As for parameters and latent variables of response time model, other than θ,

analytical expressions of full conditionals need to be modified with the definitions of $L(x)$. Please see the expressions in appendix A for more details. For full conditionals of θ please see step 2.1 in appendix A. It is worth noting that similar to DIR models as in Wang et al. (2013), forward filtering and backward sampling steps can be developed and are provided in appendix A for simulations from this full conditional distribution. it is noted that whole implementation for monotone linkage takes 25% the time it takes for IU shaped model. So there is a significant gain in implementation time.

3.3.2 Performance of DIC_p

As mentioned in earlier section we present the simulation results here. In Table 3.4, we first summarize how DIC_p is reported and misclassification rate is computed and in Table 3.5 we further summarize by reporting the misclassification rates only for different values of β and for different data models. In Table 3.4 *iter* column specifies different seed combinations for which the data can be reproduced. DIC_p *for Monotone* computes the DIC_p when we fit monotone linkage model on the data. Similarly DIC_p *for IU* computes DIC_p when IU model is fitted. Eventually whichever DIC_p is smaller is chosen as a better fit, which if matches with the true model produces a 'Correct' value for *Decision* In table 3.3.2 summary of misclassification rates is given. misclassification rate is simply defined as % of times DIC_p makes a wrong decision for different values of β when true data models are pre-specified.

It is evident from simulation study that DIC_p performs pretty well in distinguishing

β	Iter (seeds)	True Model	DIC_p for Monotone	DIC_p for IU	Decision
-0.17	100-50	IU	5857.2	5852.9	Correct
-2	90-1100	Mon	5927.5	5929.5	Correct
⋮	⋮	⋮	⋮	⋮	⋮

Table 3.4: Summary of reporting DIC_p

β	True Model	Misclassification rate
-1	IU	10%
	Mon	10%
-2	IU	10%
	Mon	10%
-10	IU	10%
	Mon	10%

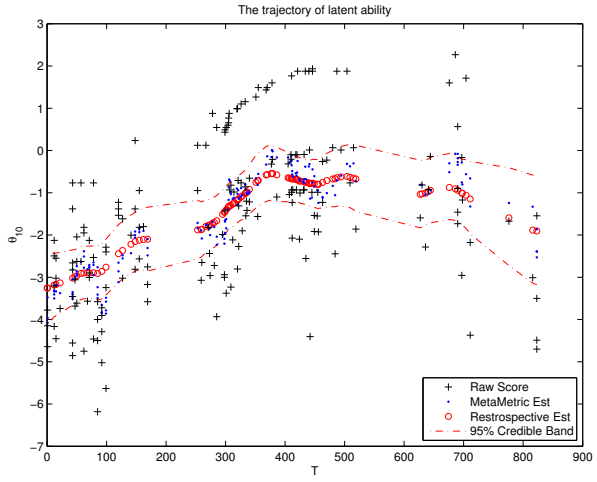
Table 3.5: Misclassification rates

the models when true data generating process is one of the models. It has been observed (but not included here) that if β is chosen a number very close to 0 I-U model gets preferred always. That is expected because as the value of β decreases distinguishing power of β decreases too.

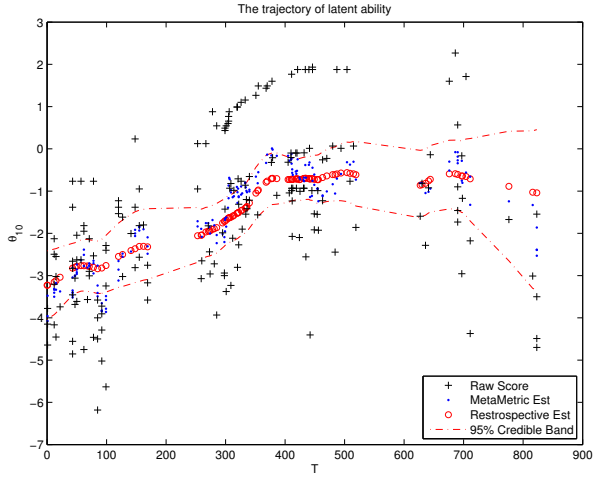
3.4 I-U vs Monotone Linkage: MetaMetrics Test Data

To the best of our knowledge, there has not been any empirical WORK for determining which among the two linkages $L(\cdot)$ indeed fits the data better for conjointly modeling response times and item responses, especially when the testing data are collected at irregular and individual varying time points for a series of computerized (adaptive)

testing. This lack of research motivates us to conduct an empirical study of MetaMetrics data-sets for identifying a better linkage. Based on the partial DIC, the inverted U-shape linkage turns out to fit the MetaMetrics data better, where the DIC_p for inverted U-shape is 5661.4 in comparison to that of monotone linkage, which is 5775.3.



(a) Monotone Linkage



(b) Inverted U-shaped Linkage

Figure 3.7: The posterior summary of the ability growth of θ_{10} for two linkages, where red circles, black plus and blue dots represent posterior median estimates of the ability, raw score and MetaMetric estimates, respectively and red dash lines represent 95% CBs.

Figure 3.7 illustrates the ability trajectories using monotone (left) and inverted-U shape (right) linkage side by side, where red dots presents the posterior median, red dash lines correspond to their credible bands (CB) and black plus indicates the 'raw score', which is a rough estimate of one's ability obtained by solving the equation that the expectation of expected score for a person's ability is equivalent to the observed score. Observed for DIR-RT models is that the estimates for the underlying increasing trend of ability growth are comparatively more robust in case of inverted-U shape than that of monotone linkage. This phenomenon is shown, for example, in the estimation of ability for θ_{10} in Figure 3.7, during the period of 350-500 days and 700-800 days, where the ability trajectory using inverted U-shape is more reluctant to change its increasing trend unless there is strong support from data (seen from raw scores (black plus) in Figure 3.7, where it is computed on the same scale as the $\theta_{i,t}$ and can be regarded as the raw data) .

As outlined in section 2.5.1 Lindley's method suggests $\beta = 0$ can not be rejected at $\alpha = 1\%$ for monotone linkage since 99% CI of β under monotone linkage includes 0, while $\beta = 0$ is rejected at both $\alpha = 1\%$ and $\alpha = 5\%$ for inverted U-shape (see Table 3.6). This indicates that the monotone linkage has weaker correlation with response times than the inverted U-shape linkage for MetaMetric datasets. Both Lindley's Method and partial DIC criterion support the choice of inverted U-shape linkage for the analysis of our proposed DIR-RT models.

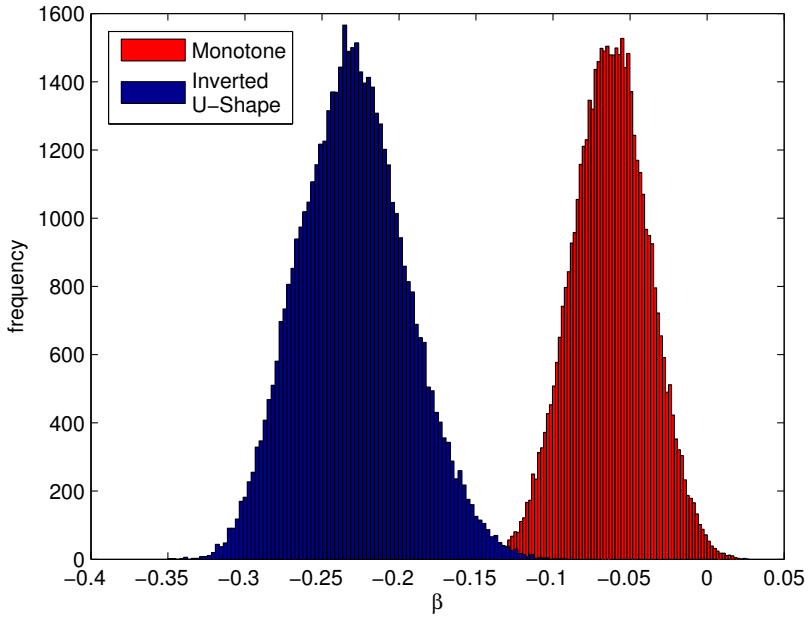


Figure 3.8: Two histograms for two linkages; I-U shaped(left), Monotone(right)

Model(β)	Inverted U-shape	Monotone
PM	-0.2305	-0.0627
95% CI	$(-0.2940, -0.1571)$	$(-0.1125, -0.0137)$
99% CI	$(-0.3105, -0.1345)$	$(-0.1317, 0.0003)$

Table 3.6: Posterior summary of β under two models, where 'PM' in the table is the abbreviation for 'posterior median'.

3.5 Discussion

In longitudinal item response theory, we presented novel modeling framework in which model comparison is executed to propose best response time linkage for education data, which allows pupils to come at irregularly spaced time-points to take multiple tests. We note that computation is very straightforward and easy as response time models are usually log-normal and reduces the computation significantly compared to other types

of DIC. This proposed partial DIC performed very well in simulated data. This DIC should work well as long as response times given ability ($\boldsymbol{R} \mid \boldsymbol{\theta}$) is independent to the response given ability ($X \mid \boldsymbol{\theta}$) and $\boldsymbol{\theta}$ is estimated well. This partial DIC may fall apart if $\boldsymbol{\theta}$ is not estimated well or $\boldsymbol{\theta}$'s system equations are mis-specified or when both occur. Also it only addresses response-time models sharing the same evolution process of $\boldsymbol{\theta}$. So if two DIR-RT frameworks have different evolution processes it may fall apart. More detailed simulation study is required for future work to study how robust it is under misspecification of the evolution process of $\boldsymbol{\theta}$. Partila DIC can not be used to compare DIR or DIR-type models as response time information is a must.

Chapter 4

Bayesian Estimation of Monotonic Ability Growth through Regularized Splines

4.1 Introduction

4.1.1 Background and Motivation

Item Response Theory (IRT) models, also called latent trait (analysis) models originated from analyzing dichotomous items (Lord (1953), Rasch (1961)). Their applications allow researchers to separate assessment of the latent ability of examinees (e.g., attitude, proficiency, preferences and other mental/behavior properties) from effectiveness of the test items. Please see section 1.1 to know more on IRT models and its origination. The large and complex data, collected from computer-based measurement testing instead of traditional paper based testing, usually have three major features, i.e., *longitudinal observations*, *local dependent responses* and *randomized items*. These make the classic

IRT models face great challenges. Please see into section 1.2.2 to know more in details about current models that address the issues. In this Chapter we are particularly interested in various growth models that were considered for ability growth process. For a literature on parametric models for ability growth, that have been implemented, please review section 1.3.2. As far as growth models were concerned, Andersen (1985), Embretson (1991) and Davier, Xu, and Carstensen (2011) treated longitudinal ability as multivariate ability vector having common multivariate distribution structure. Ability thereby loses its time-series interpretation. Later structural equations modeling (SEM) has become a popular tool to analyze linear or polynomial function of time (Hsieh et al. (2013)) with random coefficients. They are also popularly known as latent growth curve (LGC) models. Bollen and Curran (2004) made a comparative study and showed that an autoregressive latent model (ALT) performs better than latent models or AR methods. This motivated Wang et al. (2013) to combine the two ideas to propose a Markov process with time dependent co-variates . In summary, all these models are yet based on some parametric assumption; hence they may be restrictive. Although there have been prior works on non-parametric item response function (IRF), that connect ability and probability of correct response, and, there have been research works that take a functional data viewpoint on probability function bypassing the interpretation of latent ability, there have been hardly any known work to the best of my knowledge on the non-parametric smooth ability growth curves. All these models fail to incorporate an overall monotonic trend directly in the modeling framework as ability is often deemed to have

an overall monotonic growth. Also estimates from these models turn out to be zigzagged specially in the case of tests executed at irregularly spaced time points. Yet we all can safely assume that a person's growth is very smooth process for most of the cases. This necessitates innovative modeling of ability that ensures smoothness. In addition, DIR or DIR-RT type models solely build on theoretical justification of the evolution process of ability. There can be cases where this justification may not be exactly correct.

In this paper, we propose a novel class of DIR models with semi-parametric smooth growth curve (DIR-SMSG) based on certain type splines so as to make latent ability growth more flexible, thereby easy to interpret in the scenarios of multiple time test-takers at individually-varying and irregular-spaced time points. Our proposed solution is based on regularized splines. For the sake of being self-contained we shall give quick overview of splines and some of its interesting properties.

4.1.2 B-spline Functions

Definition

A spline function of order $p, f(t)$, can be defined as piece-wise polynomial where pieces are based on knots. Since our eventual goal is to work with a spline over a finite interval , say $[a, b]$, $a, b \in \mathcal{R}$, Let us introduce some notations based on this interval. Let the knots be given by $a = \zeta_0 < \zeta_1 < \cdots < \zeta_K < \zeta_{K+1} = b$ so that there are K many distinct internal knots. The knot sequence is augmented by adding $(p-1)$ replicates of

the end-points on both sides. As a result, full sequence, $\eta_1, \eta_2, \cdots, \eta_{K+2p}$ can be given in the following way

$$\eta_1 = \eta_2 = \cdots = \eta_p = \zeta_0 = a, \tag{4.1}$$

$$\eta_{p+1} = \zeta_1, \eta_{p+2} = \zeta_2, \cdots, \eta_{p+k} = \zeta_k,$$

$$\eta_{k+p+1} = \eta_{k+p+2} = \cdots = \eta_{k+2p} = \zeta_{k+1} = b.$$

The spline function, $f(t)$, is a polynomial of order p on every interval (η_j, η_{j+1}) and has $(p-2)$ continuous derivatives on the interval (a, b). The set of spline functions of order p for a fixed sequence of knots, $\boldsymbol{\eta} = \eta_1, \cdots, \eta_{K+2p}$, forms a vector space of functions. A very interesting basis for that is what are known as Basis splines or simply B-splines of size $(K+p)$ (See De Boor (2001), Curry and Schoenberg (1988)). They can be defined recursively as follows,

$$\begin{aligned} B_{j,1}(t) &= \begin{cases} 1 & \text{if } \eta_j \leq t < \eta_{j+1} \\ 0 & \text{otherwise} \end{cases} \\ B_{j,l}(t) &= \frac{t-\eta_j}{\eta_{j+l-1}-\eta_j} B_{j,l-1}(t) + \frac{\eta_{j+l}-t}{\eta_{j+l}-\eta_{j+1}} B_{j+1,l-1}(t), \end{aligned} \tag{4.2}$$

where $l = 2, \cdots, p$ and $j = 1, \cdots, K+2p-l$. If we adopt the convention that $B_{j,1}(t) = 0 \ \forall \ t, t \in \mathcal{R}$ if $\eta_j = \eta_{j+1}$, then by induction $B_{j,l}(t) = 0$ if $\eta_j = \eta_{j+1} =$. Hence, $B_{1,l}(t) =$

$0 \; \forall \; t \in \mathcal{R}$ and $l < p$ on the defined knot sequence. $f(t)$ can be expressed as

$$f(t) = \sum_{j=1}^{K+p} \beta_j B_{j,p} \, ,$$

where β_j 's are usually called control points.

Some Properties of B-splines

Now suppose one would like to estimate smooth curve by spline of certain order and with fixed knot-sequence. If one has enough data points to estimate that smooth curve then spline estimate would be nothing but the curve with control points obtained through least square. Here we note that if we increase the knot sequence and/or degree we can have larger basis to capture more complex patterns but with the increase in the number of B-splines estimates may not be smooth enough due to variance-bias trade-off. To impose smoothness on spline estimates either one has to restrict the number of B-splines or one needs to introduce regularization in some way. One interesting property of spline is that its derivative is also a spline with lower order. Since smoothness is often measured through derivatives, this property particularly has very special significance. Prochazkova (2005) showed that

$$\begin{aligned} f'(t) &= \sum_{j=1}^{k+p} \beta_j [\frac{p-1}{\eta_{j+p-1} - \eta_j} B_{j,p-1} - \frac{p-1}{\eta_{j+p} - \eta_{j+1}} B_{j+1,p-1}] \\ &= \sum_{j=1}^{k+p-1} \frac{p-1}{\eta_{j+p} - \eta_{j+1}} (\beta_{j+1} - \beta_j) B_{j+1,p-1}. \end{aligned} \tag{4.3}$$

First we note that from (4.3) increasing splines can be obtained by ensuring $\Delta\beta_j$ to be positive to make sure overall derivative function is positive at all t. This property will be made use of in the next section to impose monotonicity restriction. Also it immediately follows that l^{th} derivative of $f(t)$ would depend on values of $\Delta^l\beta_j$, l^{th} order difference of β_j's. This tells us one can increase smoothness of splines by shrinking these higher order differences of β_j's

P-spines or Penalized Splines

P-splines or penalized splines are nothing but regularized splines. The objective is to find a smooth spline estimate of a given mean function that may appear withing modeling framework. As shown in section 4.1.2, Eilers and Marx (1996) applied this idea to penalize higher orders to get regualarized estimates of β_j's. So in frequentists' sense it was maximization of the following :

$$l(\beta, \tau, \psi|\text{data}) - \lambda \sum_{k\geq 2} (\Delta^k \beta_j)^2. \tag{4.4}$$

Eilers and Marx (1996) only chose to penalize 2nd order difference for the sake of simplicity. Later Brezger and Lang (2006) generalized this idea to Bayesian framework. Our smooth estimates of ability are motivated by Brezger and Lang (2006) 's idea.

4.1.3 Preview

In section 2, we will put forward a new class of dynamic item response models with semi-parametric and smooth ability growth (DIR-SMSG). Due to complexity of the model considered, Bayesian methods and Markov Chain Monte Carlo (MCMC) computational techniques will be employed and section 3 will present the statistical inference procedures. Section 4 validates Bayesian inference procedure proposed with some simulations and this section will study the robustness of the model while fitted in the simulated data, that originated from DIR models. We also explore the possibility of applying it to MetaMetrics data. In section 6, we conclude by stressing the robust recovery of the parameter estimates of response models as well as smooth monotonic estimation of ability growth.

4.2 Dynamic Item Response with Semi-parametric Smooth Growth (DIR-SMSG)

Motivated by DIR models as introduced by Wang et al. (2013) we propose a two-stage model. In the first stage response is modeled through one parameter logistic model as in DIR models with latent ability level. This can be extended to 2 parameter logistic (2-PL) or 3 parameter logistic (3-PL) if required. In the 2nd stage the dynamic latent growth is captured via semi-parametric methods using splines. Later the monotonicity and

smoothness property is ensured with appropriate choices of priors for spline parameters.

4.2.1 First Stage: The Observation Equations in DIR-SMSG Models

To express the models, we shall borrow similar notations as introduced in section 2.2.1. Let $X_{i,t,s,l}$ be the item response to indicate the correctness of the answer of the l-th item in the s-th test on the t-th day given by the i-th person, where $i = 1, \cdots, n$ (number of subjects); $t = 1, \cdots, T_i$ (number of test dates); $s = 1, \cdots, S_{i,t}$ (number of tests in a day); and $l = 1, \cdots, K_{i,t,s}$ (number of items in a test). Likewise, denote the difficulty of the l-th item as $d_{i,t,s,l}$. Here we shall need to make a clarification of the notation. Let t be a point on a time scale used for all individuals. It should be clear that if i^{th} individual takes exams on T_i many test-dates. In general, tests taken on h^{th} day since beginning may not be the same for every individual as exact time point, t_h is nested within i^{th} individual. For the sake simplicity we shall often use $X_{i,t,s,l}$ to mean $X_{i,t_i,s_{t_i},l_{s_{t_i}}}$

The Observation Equations of Item Responses

Observation equation remains the same as in DIR models. We shall briefly re-visit section 2.2.1. Let us recall in a design of computerized tests, item difficulty, i.e., $d_{i,t,s,l}$, is a randomized parameter, assuming to be randomly drawn from a bank of items with certain ensemble mean. $d_{i,t,s,l}$ then can be modeled as a measurement error model, where $d_{i,t,s,l} = a_{i,t,s} + \epsilon_{i,t,s,l}$ with $a_{i,t,s}$ being an ensemble mean difficulty of items in the s-th test,

and $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with σ^2 known according to the test design and $\mathcal{N}(\cdot, \cdot)$ denoting a normal distribution. Similar as Wang et al. (2013) did, we modify classic IRT models to accommodate the complication by modeling the observation equation of item responses as

$$\Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}) = \mathrm{F}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}), \tag{4.5}$$

where $\theta_{i,t}$ represents the i-th person's ability on the day t with assuming one's ability is constant over a given day, $\varphi_{i,t}$ and $\eta_{i,t,s}$ take account of daily and test random effects, respectively, to explain the possible local dependence of item responses. Assume $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$ with its precision unknown and being different for each person. Similarly, let $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$ with $\eta_{i,t} = (\eta_{i,t,1}, \ldots, \eta_{i,t,S_{i,t}})'$ being the vector of test random effects on the day t for the individual i and $\mathbf{I}$ is an $S_{i,t} \times S_{i,t}$ identity matrix. Utilizing precision parameters in place of variance parameters for normal distributions is because of the convenience in Bayesian computation. The reason of letting $\eta_{i,t}$ be a singular multivariate normal (by setting the sum of test random effects to be zero on any day t) is to remove any possibility of unidentifiable issues between daily and test random effects. In the application to MetaMetrics testbed, choose $F(\cdot)$ to be a logistic link due to the convention in MetaMetrics, where they used logit unit as a linear transformation of Lexile scale used in their products.

4.2.2 Second Stage: System Equations in DIR-SMSG

In this stage we deviate from DIR models significantly. Instead of choosing to model as a parametric Markov chain we choose to treat the ability as an unknown monotonic and smooth function of time. Usually it is assumed that the true inherent mean ability growth is monotonic and smooth, thus allowing some variability for realized ability process. The reason of introducing this variability is simply the fact though students' ability usually shows a monotonic growth there are patches of time, when people may not always realize their full potential or they are super-prepared for the tests. Thus $\theta_{i,t}$ can be given by the following

$$\begin{aligned} \theta_{i,t} &= \mathcal{F}_i(t) + w_{i,t}, \quad w_{i,t} \sim \mathcal{N}(0, \Delta_{i,t}/\phi) \ \forall \ t \geq 2, \\ \mathcal{F}_i(t) &\uparrow \ t. \end{aligned} \tag{4.6}$$

The first term $\mathcal{F}_i(t)$ is assumed to be unknown but fixed function of time where as $w_{i,t}, \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, represents the random component of the change in the i-th person's ability on the t-th day with ϕ being a common known parameter. In reality ϕ may or may not be known. For the sake simplicity we assumed ϕ to be known. This assumption of $w_{i,t}$ presumes that one's ability is much more uncertain, if he/she is absent for a longer period.

In general $\mathcal{F}_i(t)$ can be very hard to estimate along with monotonicity and smoothness restriction and may require lots of data points. To circumvent this hurdle we further

assume that $\mathcal{F}_i(t)$ can be approximated by a spline function (De Boor (2001)) of order 4 with equi-distant knot sequence as follows

$$\mathcal{F}_i(t) = \sum_{j=1}^{m} \alpha_{i,j} B_j(t_i), \quad \alpha'_{i,j} s \uparrow \text{w.r.t j}, \tag{4.7}$$

where m represents size of B-spline basis and $\alpha_{i,j}$'s are control points. m=p+K, where p denotes the order of the spline (such as 3 for quadratic, 4 for cubic splines) and K denotes the total number of equi-spaced internal knots over whole time span of growth (please review section 4.1.2 for more details). Choosing equi-spaced knots helps computations and does not pose any problem as long as there are enough data points to estimate the control points. In both simulated data and real-data sets we will be working with cubic splines and 20(m) many b-splines. $\alpha_{i,j}$'s are the control points for i-th individual and they increase with j. This property ensures monotonicity of spline function.

4.2.3 A Summary of DIR-SMSG Models

To summarize, the proposed one-parameter DIR-SMSG models have two-stages, the 1st stage consists of observation equations, while the 2nd stage consists of system equations,

$$\begin{aligned}
&\textit{1st stage:} \qquad \Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}, \epsilon_{i,t,s,l}) \\
&\qquad\qquad = \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}, \\
&\textit{2nd stage:} \qquad \theta_{i,t} = \sum_{j=1}^{m} \alpha_{i,j} B_j(t_i) + w_{i,t}, \quad \alpha'_{i,j} s \uparrow \text{w.r.t j},
\end{aligned}$$

where $X_{i,t,s,l}$ is observed; $a_{i,t,s}$'s and $\Delta_{i,t}$'s are known and $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with known σ^2. Moreover, we have the following distribution assumptions. $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, ϕ is also assumed to be known.

4.3 Statistical Inference and Bayesian Methodology

As discussed in the section 1.5 we note that a frequentist's approach implementing Expectation Maximization (EM) or some version of that or marginalized maximized likelihood estimators (MML) is almost next to impossible to compute due to extremely complex nature of likelihood. In addition, some parameter spaces ($\alpha'_{i,j}s$) are restricted, which renders the EM or MML method extremely intractable. On top of that standard error estimates of the estimators are not very reliable. On the other hand Bayesian methodology not only simplifies modeling and estimating the uncertainties, thanks to advancements in MCMC techniques, Bayesian computation is way simpler and easy to be extendable to other complex variants of the model. Next we describe and implement a fully Bayesian methodology.

4.3.1 Prior Distribution for the Unknown Parameters

For the prior choices of scale parameters δ_i's, τ_i's and ϕ, we use the same choices of Wang et al. (2013), which assign them objective priors, $\pi(\phi) \propto 1/\phi^{3/2}$, $\pi(\delta_i) \propto 1/\delta_i^{3/2}$, and $\pi(\tau_i) \propto 1/\tau_i^{3/2}$ for all i.

In this study we plan to use 20 B-splines of order 4 (cubic splines) that might lead to over-fitting. As a result finally estimated spline curve may not be smooth. To avoid that, we would like to use the idea of implementing 2nd order smoothness penalty as proposed originally by Eilers and Marx (1996). In our case, following the idea of Brezger and Steiner (2008), we will be replacing the frequentist's notion of 2nd degree difference penalization by its stochastic analogue of 2nd order random walk as prior for spline coefficients (i.e. $\alpha_{i,j}$'s) in Bayesian framework.

$$\begin{aligned} \alpha_{i,j} &= 2\alpha_{i,j-1} - \alpha_{i,j-2} + u_{i,j}, u_{i,j} \sim \mathcal{N}(0, \omega_i^{-1}), \quad \pi(\alpha_{i,1}), \pi(\alpha_{i,2}) \propto 1 \quad (4.8) \\ \text{or} \quad D^2\boldsymbol{\alpha}_i &\sim N(\mathbf{0}, \omega_i^{-1} I_{m-2}). \end{aligned}$$

In the above equation ,the operator D^2 is the following matrix,

$$\begin{bmatrix} 1 & -2 & 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 1 & -2 & 1 \end{bmatrix}.$$

In this formulation ω_i's are hyper parameters that control the degree of smoothness of individual specific mean ability growth curve. For example higher values of ω_i's would ensure smoother $\alpha_{i,j}$'s values, which in turn ensures smoother spline estimates.

Incorporating monotonicity restriction via prior for $\alpha_{i,j}$ can be given as follows:

$$\begin{aligned} p(\boldsymbol{\alpha}_i|\omega_i) \propto \quad & \exp(-\omega_i \boldsymbol{\alpha}_i{}' D^{2'} D^2 \boldsymbol{\alpha}_i/2) \prod_{j\geq 2} 1_{\alpha_{i,j}\geq\alpha_{i,j-1}}, \\ \text{or} \quad & \exp(-\omega_i \boldsymbol{\alpha}_i{}' K^{\delta} \boldsymbol{\alpha}_i/2) \prod_{j\geq 2} 1_{\alpha_{i,j}\geq\alpha_{i,j-1}}. \end{aligned} \tag{4.9}$$

The prior for ω_i is assumed diffused Gamma, a weakly informative prior.

4.3.2 Posterior Distribution and Data Augmentation Scheme

Using the fact that a standard logistic distribution can be expressed as a scale mixture of normals and applying the data augmentation idea as employed in section 2.3.2 (Tanner and Wong (1987)), a latent variable $Y_{i,t,s,l}$ can be introduced for each response variable $X_{i,t,s,l}$, where $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}, 4\nu^2_{i,t,s,l})$ and $\Pr(X_{i,t,s,l} = 1|\theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) = \mathrm{P}(Y_{i,t,s,l} > 0|\theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l})$. Let us define $X_{i,t,s,l} = 1$ if $Y_{i,t,s,l} > 0$ and $X_{i,t,s,l} = 0$ otherwise, and the introduction of $Y_{i,t,s,l}$ can facilitate the MCMC computation although it introduces more unknowns. Since $\epsilon_{i,t,s,l} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, marginalizing out it, results in $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, 4\nu^2_{i,t,s,l} + \sigma^2)$.

Then, the one-parameter DIR-SMSG can be rewritten as

$$\begin{aligned} \theta_{i,t} &= \sum_{j=1}^{m} \alpha_{i,j} B_j(t_i) + w_{i,t}, \\ Y_{i,t,s,l} &= \theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \xi_{i,t,s,l}, \\ X_{i,t,s,l} &= I(Y_{i,t,s,l} > 0), \end{aligned} \tag{4.10}$$

where $\xi_{i,t,s,l} \sim \mathcal{N}(0, \psi^{-1}_{i,t,s,l})$ with $\psi^{-1}_{i,t,s,l} = 4\gamma^2_{i,t,s,l} + \sigma^2$ and $\gamma_{i,t,s,l} \sim$ K-S distribution, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$. If we treat ϕ known in the above expression, one can marginalize $\boldsymbol{\theta}$ out to simplify the models, which will in turn help simplify MCMC computation steps in two ways. First dropping θ would reduce the number conditionals to loop through and secondly convergence would be faster due to collapsing. The equations in 4.10 can be re-written as follows,

$$\begin{aligned} Y_{i,t,s,l} &= \sum_{j=1}^{m} \alpha_{i,j} B_j(t_i) - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \xi^*_{i,t,s,l}, \\ X_{i,t,s,l} &= I(Y_{i,t,s,l} > 0), \end{aligned} \tag{4.11}$$

where $\xi^*_{i,t,s,l} \sim \mathcal{N}(0, \psi^{-1}_{i,t,s,l} + \Delta_{i,t}/\phi)$ with $\psi^{-1}_{i,t,s,l} = 4\gamma^2_{i,t,s,l} + \sigma^2$ and $\gamma_{i,t,s,l} \sim$ K-S distribution, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$.

Define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n)'$ with $\boldsymbol{\theta}_i = (\theta_{i,0}, \theta_{i,1}, \cdots, \theta_{i,T_i})'$; $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n)'$ with $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,m})'$; $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_n)', \boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)', \boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)'$; $\mathbf{Y} = \{Y_{i,t,s,l}\}, \boldsymbol{\gamma} = \{\gamma_{i,t,s,l}\}$ and $\mathbf{X} = \{X_{i,t,s,l}\}$; $\boldsymbol{\varphi} = \{\varphi_{i,t}\}$, $\boldsymbol{\eta} = \{\eta_{i,t,s}\}$ and $\boldsymbol{\eta}^*_{i,t} =$

$(\eta_{i,t,1}, \cdots, \eta_{i,t,S_{i,t}-1})'$; where $l = 1, \cdots, K_{i,t,s}, s = 1, \cdots, S_{i,t}, t = 1, \cdots, T_i$ and $i = 1, \cdots, n$. Let us also introduce here $T_i \times m$ dimensional matrices, X_i, defined as (($X_i(t_i, j) = B_j(t_i)$)). Please note that X_i is not boldfaced to be distinguished from data $\mathbf{X}$. Let bold symbols denote the vector on the omitted subscripts.(e.g. $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \cdots, \alpha_{i,m})^t$).Given the data $\mathbf{X}$, the joint posterior density of $(\mathbf{Y}, \boldsymbol{\theta}, \phi, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\alpha})$ of our proposed DIR-SMSG models is

$$
\begin{aligned}
&\pi(\mathbf{Y}, \boldsymbol{\theta}, \phi, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma} \mid \mathbf{X}) \\
&\propto \{p(\phi) \prod_{i=1}^{n} p(\tau_i) p(\delta_i) p(\boldsymbol{\alpha}_i|\omega_i) p(\omega_i|a,b)\} \{\prod_{i=1}^{n} \prod_{h=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} p(\gamma_{i,t,s,l})\} \\
&\times \{\prod_{i=1}^{n} \prod_{h=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} [I(Y_{i,t,s,l} > 0) I(X_{i,t,s,l} = 1) + I(Y_{i,t,s,l} <= 0) I(X_{i,t,s,l} = 0)] \\
&\sqrt{\frac{\psi_{i,t,s,l}}{2\pi}} \exp(-\frac{\psi_{i,t,s,l}}{2}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2) \times I(\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s})\} \\
&\times \{\prod_{i=1}^{n} \prod_{h=1}^{T_i} \tau_i^{\frac{S_{i,t}-1}{2}} \exp(\frac{-\tau_i \eta_{i,t}^{*\prime} \Sigma_{i,t}^{-1} \eta_{i,t}^{*\prime}}{2})\} \times \{\prod_{i=1}^{n} \prod_{h=1}^{T_i} \delta_i^{\frac{1}{2}} \exp(\frac{-\delta_i \varphi_{i,t}^2}{2})\} \\
&\times \{(\prod_{i=1}^{n} \prod_{h=1}^{T_i} \phi^{1/2} \exp(-\phi \frac{(\theta_{i,t} - X_i(t,:)\boldsymbol{\alpha}_i)'(\theta_{i,t} - X_i(t,:)\boldsymbol{\alpha}_i)}{2\Delta_{i,t}})\}.
\end{aligned}
$$

The proof of posterior propriety of DIR-SMSG models closely follows from a simple extension of Appendix C in Wang et al. (2013) for DIR models. The expression given above works for a general case when ϕ is not known. If ϕ is known the expression would remain the same except for few changes like $p(\phi)$ is omitted and θ needs to be re-defined as in Appendix C. Please see Appendix C for the expression for posterior density when

ϕ is known.

4.3.3 MCMC Computation of DIR-SMSG Models

The computation is carried out by MCMC scheme that samples from the posterior via block Gibbs sampling schemes. The difficulty of the sampling scheme is to draw the spline control points, $\alpha_{i,j}$'s. It turns out that the full conditional distribution of $\boldsymbol{\alpha}_i$ is a truncated multivariate normal distribution. One can get a block sample by running short MCMC chain (usually becomes stable after 100 iterations) from this full conditional (by Robert (1995)'s approach)or one can consider one run from each univariate truncated normal and integrate that into overall Gibbs sampler. We adopted the first approach in which every time we need to draw a sample from full conditionals of α we ran a sub-chain of length 100 before moving to the next draw from other full conditionals of the gibbs sampler. The details of MCMC steps are given in Appendix C. The Gibbs sampling starts at *Step 1* in Appendix C, with initial values for $\theta^{(0)}$, $\phi^{(0)}$, $\varphi^{(0)}$, $\eta^{(0)}$, $\delta^{(0)}$, $\tau^{(0)}$, $\omega^{(0)}$,$\gamma^{(0)}$ and $\alpha^{(0)}$ then loops through *Step 10* in Appendix C, until the MCMC converges. The initial values chosen in the applications were $\theta^{(0)} = \vec{0}$, $\phi^{(0)} = 1$, $\varphi^{(0)} = \vec{0}$, $\eta^{(0)} = \vec{0}$, $\delta^{(0)} = \vec{1}$, $\tau^{(0)} = \vec{1}$, $\gamma^{(0)} = \vec{1}$, $\omega^{(0)} = 0.5\vec{1}$. Here we note that the initial value of $\alpha^{(0)}$ can not be assigned constant vector because of monotonicity restriction. Let's define a monotonic sequence as follows

$$\alpha^{(0)}[i,j] = 0.5 + (0.002(i-1)) + (0.05i)(j-1), \forall\, i,j; i = 1 \cdots n; j = 1 \cdots m.$$

The method described here is for the general case when ϕ may not be known. In our case ϕ is known. (ϕ may be known from prior studies or can be roughly elicited from the raw estimates). Known ϕ would simplify and accelerate the MCMC steps. The modified steps are also given in Appendix C. The convergence was evaluated informally by looking at trace plots of all the parameters and ability curves. Then, statistical inferences are made straightforward from the MCMC samples. For example, an estimate and 95% credible interval (CI) for the latent trajectory of one's ability $\theta_{i,t}$ can be plot from the median, 2.5%, and 97.5% empirical quantiles of the corresponding MCMC realizations. In these examples, ability will again be graphed as a function of time,t, so that the dynamic changes of an examinee is apparent.

4.4 Simulation Study

To validate the inference procedure and compare the benefits of a semi-parametric set-up , a simulation study was conducted with similar set-up as that of DIR models as laid out in section 4 of Wang et al. (2013). To save the space, we only illustrate the situation when data generating process follows an monotonically increasing mean ability process. The simulation method considers multiple individuals taking a series of tests scheduled at individually-varying and irregularly-spaced time points.

4.4.1 DIR-SMSG Models Simulation

Following the simulation study of DIR models in Wang et al. (2013), assume there are 10 individuals, each of them has taken four tests on 50 different test dates, where each test contains 10 items. The specification means $K_{i,t,s} = 10$, for $s = 1, \cdots, S_{i,t}$, $t = 1, \cdots, T_i$, $i = 1, \cdots, n$ with $S_{i,t} = 4$, $T_i = 50$ and $n = 10$. Let time lapse between two consecutive test dates be $\Delta_{it}, = t + 10$ if $t \leq T_i/2$ or $\Delta_{i,t} = t - 10$ otherwise, creating a irregularly spaced gap between two test dates.

In order to do the comparison of DIR-SMSG models later with DIR models, we assign same values of the common parameters, ϕ, δ_i, τ_i, as used in Wang et al. (2013), where $\phi = 1/0.0218^2$, leading standard deviation of $w_{i,t}$ in the system equation (2.5) is $0.0218\sqrt{\Delta_{i,t}}$ and the values of δ_i, τ_i are specified in Table 4.7.

i	1	2	3	4	5	6	7	8	9	10
δ	2.0408	1.3333	1.8182	1.2346	1.5873	1	2.2222	1.0526	1.1494	2
τ	4	3.1250	4.3478	2.7027	3.7037	2.8571	4	2.2222	9.0909	4.5455

Table 4.7: Values of common parameters with DIR models, used in the simulation

Here ω_i is assumed to be 1 for all $i = 1 \cdots n$. Last we need to specify some values that conform prior belief about $\alpha_{i,j}$'s. In order to accomplish that, we simulate a set of values from prior process. We choose $\alpha_{1,i}$ and $\alpha_{2,i}$ quite arbitrarily such that $\alpha_{1,i} < \alpha_{2,i}$. Next we simulate $\alpha_{3,i} \cdots \alpha_{m,i}$ such that $(\alpha_{3,i}, \alpha_{4,i}, \cdots, \alpha_{m,i} | \alpha_{1,i}, \alpha_{2,i})' \sim \mathcal{N}_{\mathcal{T}}(D^{\star -1}\mathcal{C}_i, \mathcal{P}_i)$ where $\mathcal{C}_i = (2\alpha_{2,i} - \alpha_{1,i}, -\alpha_{2,i}, 0, \cdots, 0)'$ and $\mathcal{P}_i = \omega_i D^{\star\prime} D^{\star}, D^{\star} = D^2[:, 3:m]$ and truncation is based on the boundary condition $\alpha_{2,i} < \alpha_{3,i} < \cdots < \alpha_{m,i}$. We employ MCMC to

simulate $\alpha_i's$ using Robert (1995)'s approach. Steps for $\boldsymbol{\alpha}$ from *step 3*, Appendix C can be mimicked by making the following updates:

$$\text{Set } \boldsymbol{\alpha}_i^m = D^{\star -1}\mathcal{C}_i, \boldsymbol{\alpha_i^{(0)}} = \boldsymbol{\alpha}_i^{(s)}$$

$$P_i = \mathcal{P}_i,$$

where $\boldsymbol{\alpha}_i^{(s)}$ is the vector of starting points from the domain region. $\boldsymbol{\alpha}_i^{(s)}$ is chosen as $\alpha_i^{(0)}$ (see 4.3.3). We would like to point out that $D^{\star -1}$ has a very special structure as given below,

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \\ n & n-1 & \cdots & 1 \end{bmatrix}.$$

Simulation proceeds by simulating random effects or latent variables using the assigned parameter values above for DIR-SMSG models. Once we get the simulated values for $\theta_{i,t}$ using *2nd stage* model, then the test difficulties, $a_{i,t,s}$ in *1st stage* model is set to be $\theta_{i,t} + \zeta^*$, where ζ^* is a random variable with uniform distribution on $(-0.1, 0.1)$. The values of $\epsilon_{i,t,s,l}$ are drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.7333$. Notice the values of σ and ϕ are also used later for fitting simulated data as generated for DIR models in Wang et al. (2013). The dichotomous data of item responses is now treated as our observations, and the Bayesian methodology from section 4.3.3 is implemented in estimating the model

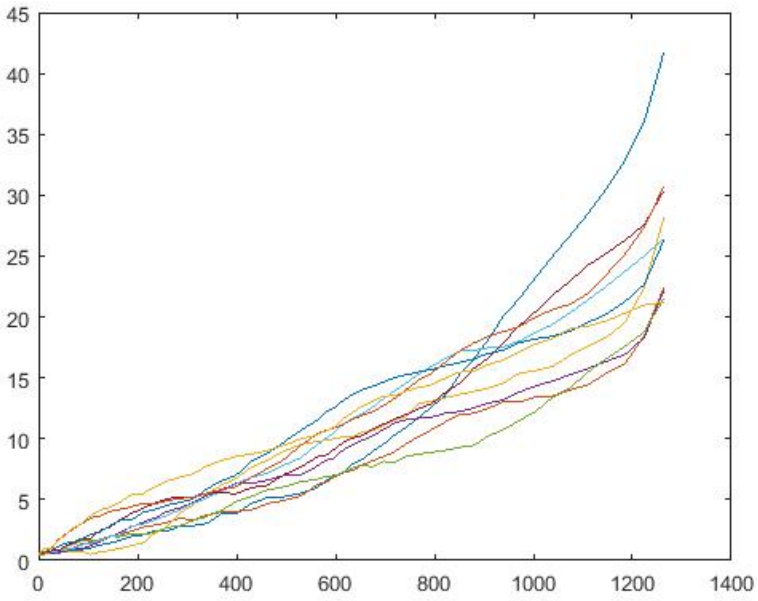


Figure 4.9: True mean ability growth curves, smooth and monotonic, based on semi-parametric model

parameters of DIR-SMSG models.

Similar to cases of DIR-RT models, the parameters are estimated through posterior median calculated from their corresponding MCMC samples. Each MCMC chain was run for 50,000 iterations with a 25,000 burn-in period. Figure 4.10 (a)-(b) give posterior median estimates (red squares) along with 95% CIs (red bars) of $\tau^{-1/2}$, $\delta^{-1/2}$ respectively and illustrate their true values (black dot). Clearly from Figure 4.10, the true values of those parameters are contained within their corresponding 95% CIs.

For the posterior median estimates of other parameters, we obtained very similar results as obtained in DIR simulation study. Median a posteriori eastimates are very

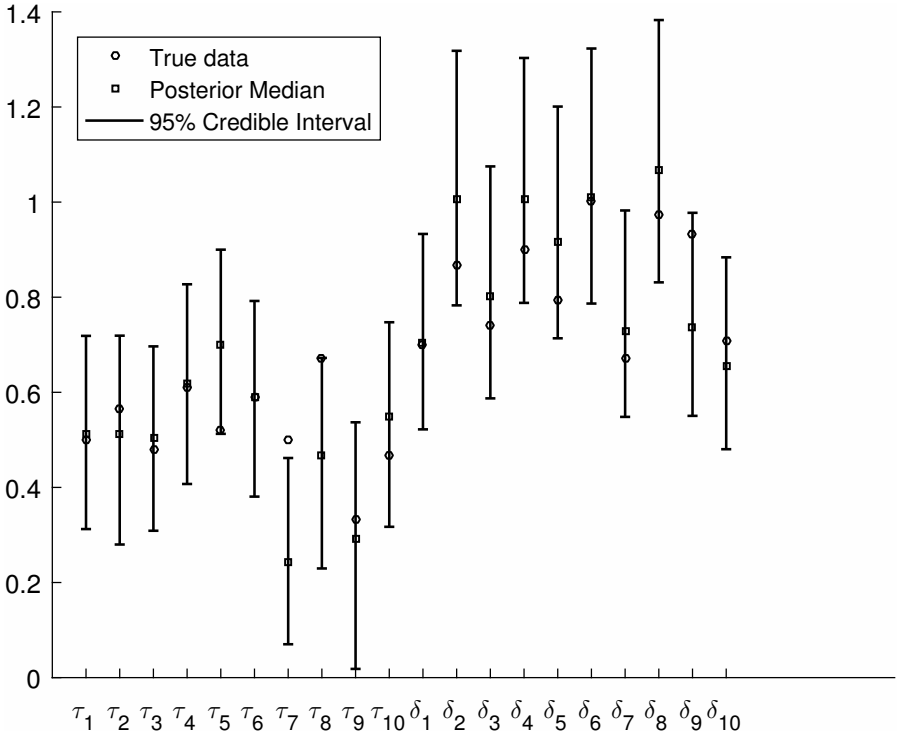
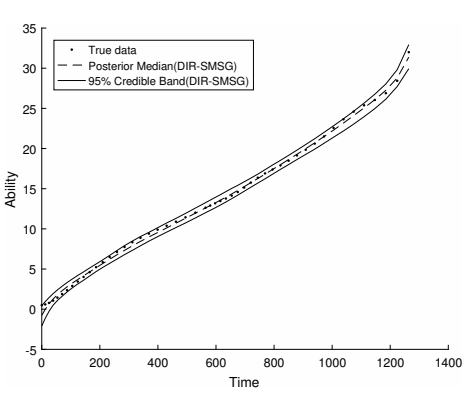


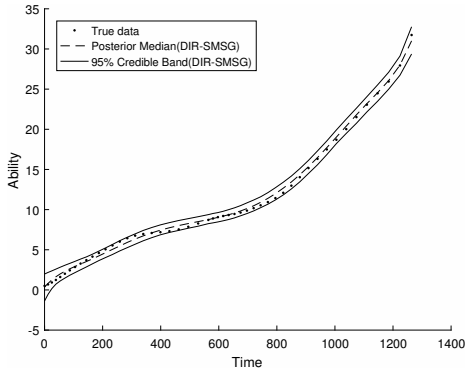
Figure 4.10: Posterior summary of $\tau_i^{-1/2}$, $\delta_i^{-1/2}$'s, where red circles represent true values, red squares are the posterior median estimates and red bars indicate 95% CIs.

close to true values and the true parameters values always have high 95% coverage probability (CP). We bring our attention to our primary interest, estimating latent ability trajectories. Figure 4.11 (a)-(d) illustrate four types of growth curves in our simulation, where (a) θ_2 represents an individual with steady increasing growth; (b) θ_6 indicates an increasing growth but it has a point of inflexion somewhere in the middle of time-line such that the growth gradually changes from slowly increasing to fast increasing. (c) θ_7 presents us with another monotonically increasing pattern with possibly two points of inflexion; (d) θ_8 displays very steady monotonic growth except for end points where growth is steeper. In Figure 4.11, the true ability curves (black dots) have been plotted

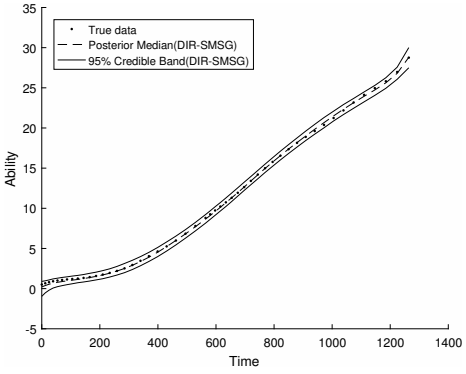
along with our posterior median estimates of ability (blue circles) and their corresponding 95% credible band (blue lines). Notice in each sub-figure, very small proportion of true values are outside of 95% credible bands. Note that the estimates of $\alpha_{i,j}$'s are not checked directly. Rather the constructed 95% credible band for monotonic mean ability growth process is indirect validation as it is equivalent to looking at MCMC estimates of $X_i(t,:)\boldsymbol{\alpha}_i$.



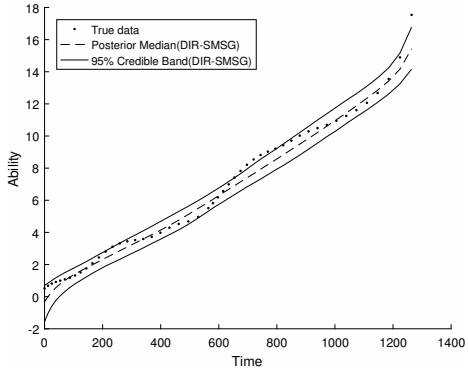
(a) 2nd individual



(b) 6th individual



(c) 7th individual



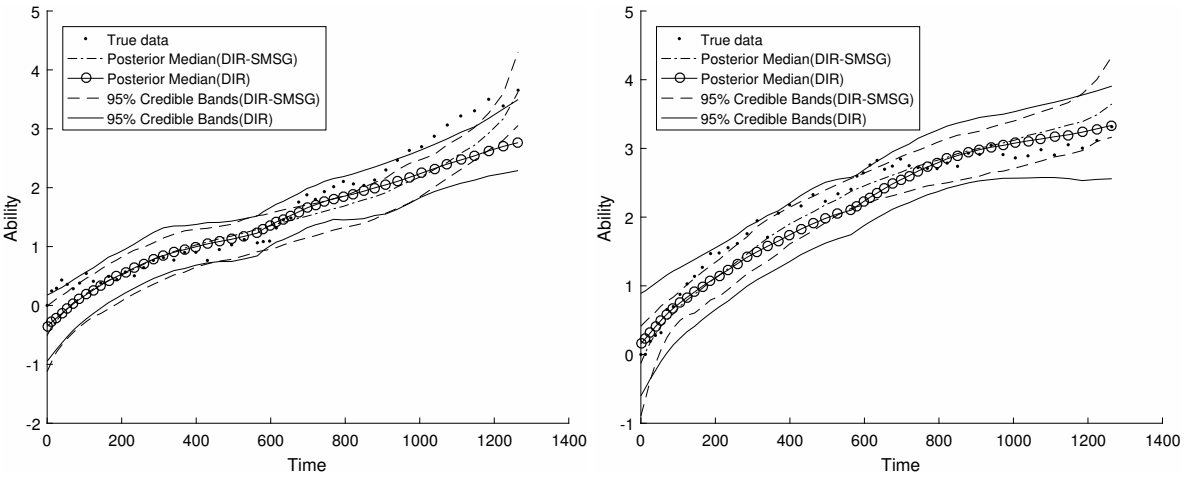
(d) 8th individual

Figure 4.11: The latent trajectory of one's ability growth, where black dots, middle dashed-line and connected lines represent true ability, the posterior median estimates and the 95% credible bands, respectively.

4.5 Robustness of DIR-SMSG

We intend to investigate the robustness of semi-parametric DIR-SMSG. In order to achieve this goal, DIR-SMSG models are fitted to the data, that was originally simulated from DIR-RT models. Trajectories of mean ability were plotted for both DIR models and DIR-SMSG models along with their corresponding credible bands. Figure 4.12 displays the growth curve of two selected individuals (i.e., θ_1 and θ_2), where the statistical inference is based on the simulated example in 2.4. For other individuals, results are similar and to save the space, we omit the plots for others. In Figure 4.12 (a) and (b), 95% CIs of DIR models (connected lines) encompass 95% CIs of DIR-SMSG models (dashed line); both 95% CIs contain the true values (black dots). The average length of 95% credible band of ability estimates for DIR-SMSG models is slightly shorter than that of DIR models. In addition, notice that in Figure 4.12,(a) both estimates (posterior median) of the graphs of DIR (circles) as well as of DIR-SMSG (dashed middle line) adhere to true ability (black dots) but in higher end points of time DIR performs poorly whereas thanks to the monotonic development DIR-SMSG tried to capture it better. The results illustrate that by incorporating information on properties like monotonicity and smoothness, we can reduce the bias of the estimates of one's ability trajectory on top of improving the precision if the true curve indeed satisfies those properties. In summary, posterior median growth curve of DIR-SMSG seems to be approximating the mean growth well and lies well within 95% credible interval of mean ability of the

posterior estimates of the DIR models. It also estimates precision parameters of response model quite well. Overall the fit is quite robust.



(a) 1st individual (b) 2nd individual

Figure 4.12: The comparison of ability estimates between DIR-SMSG and DIR models, where black dots, blue circles, middle red-dashed line represent true mean ability, DIR ability and DIR-SMSG ability estimates respectively; connected-lines (blue) and dash (red) lines represent 95% credible bands for DIR and for DIR-SMSG respectively.

4.6 Discussion

Although there have been prior works on non-parametric item response function, that connects ability and probability of correct response, and, there have been research works that take a functional data viewpoint on probability function bypassing the interpretation of latent ability, there have been hardly any known work to the best of our knowledge on the non-parametric smooth ability growth curves in the context of longitudinal IRT. The smoothness is assured by using penalized splines in Bayesian context, thereby

introducing regularization in ability curve estimation. There is no currently existing methodology that, in addition to addressing these properties of a general longitudinal testing data, generalizes the ability curve to monotone penalized splines. Our proposed Bayesian hierarchical modeling approach provides this desired unified framework that has the potential to introduce flexible interpretable models. Simulation study showed a very interesting insight that in presence of some extra information like smoothness and monotonicity, the precision and bias can be improved upon. Also it is quite robust and does not require any explicit assumption on the evolution of ability process other than curvature properties like smoothness and/or monotonicity etc. Such a framework is also readily applicable to any dichotomous psychometric tests.One of the added advantages that this semi-parametric model has over DIR or DIR-RT models is that one can get an estimate of mean ability at any given point of time within the growth period. This capability of interpolating ability is not straight-forward in DIR or DIR-RT models.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

In this Chapter we summarize the results and discuss future steps. In Chapter 2 we found out that incorporation of response time into the item response model for the analysis of individual varying and irregular spaced longitudinal observations has significantly improved precision and reduced bias for the ability estimation. Analyzing MetaMetric datasets with the help of DIR-RT models further supports findings of Wang et al. (2013). For example, the evidence of violation of local dependence assumption is generally strong in DIR-RT models, and use of test and daily random effects to model the local dependence seems to be necessary and successful; and the retrospective analysis of ability estimation is of considerable use in understanding population behavior, such as frequently observed drops in ability after a long pause in testing. DIR-RT models improve precision of other parameters of response model. The result favors the inverted U-shape linkage, which is quite important and meaningful conclusion, since it supports that for times series of testing data, the psychology of students in a test seems to be to spend more time on tests that match their ability levels and spend less time on those,

that are either too easy or too hard.

In Chapter 3 we proposed partial likelihood based DIC (DIC_p) that performs well in simulated data. When the same is applied to test which linkage is the best choice in MetaMetric testing data it re-confirmed the I-U shaped linkage, that was already proposed based on some empirical evidence in Chapter 2. In Chapter 3, semi-parametric ability growth was introduced and posterior estimation is shown to capture efficient parameter recovery. It is then shown to be quite robust as it approximates simulated data from DIR models quite well. The advantage of this estimation process is that (1) smooth and (2) monotonicity can be preserved explicitly while modeling. It also improves precision compared to DIR fits thanks to the smoothness penalty as long as ability shows monotonic growth on an average. Interpolating the ability at a given time point is very straight-forward. Finally we focused on clustering of longitudinal data, mostly with a view to cluster ability estimates based of rate of change. We proposed and applied a derivative spline based approach which captures the true clusters reasonably well in simulated set-up with some level of noise. This methodology can be readily applied to simulated data from DIR or DIR-RT models and also the MetaMetric test-data. . An alternative to this method is model-based clustering, which can be implemented through modeling c_i's by a mixture distribution etc in future works. Both the implementations are discussed in 3rd section of this Chapter.

5.2 Some Immediate Extensions

In Chapter 2, no sensitivity analysis of DIC_p is studied for mis-specified ability process. So there is some future works possible, that may add to credibility of the measure. Semi-parametric smooth growth models can be applied to MetaMetrics test data judiciously after a careful treatment of outliers as the growth may not be always monotonic for all students (This fact was immediate from Wang et al. (2013)'s study). This model also lacks the response time component and can be extended to DIR-RT type of framework and precision may be improved. Also it treats the precision parameter, ϕ known, which can be relaxed in future studies. The MCMC steps are also laid down in the Appendix C in case of precision parameter, ϕ, being unknown. All these models can be extended to 2-PL or 3-PL and normal ogive models.

5.3 Work in Progress: Clustering Ability Growth based on Rate of Learning

5.3.1 Background and Motivation

Longitudinal study in item response theory is still in its nascent stage, especially in the context of test-takers, who repeatedly come back for taking tests at irregularly spaced time points. Most of the earlier studies exploited the fact that tests are executed at equi-spaced time points, that are fixed for all test takers, for modeling IRT and

concurrently developing other important measures. In addition, clustering these ability growth trajectories is hardly formally addressed to the best of our knowledge in the context of IRT. Clustering such trajectories can have some important implications. For example, one might want to group students whose learning patterns are similar. Such cluster may not be very appealing as students studying in different grades shall have different ability intercepts at any given time point to begin with. As a result, there would be many groups and hence clustering may not serve any general interest. In contrast, one might be interested to cluster the growth of latent trait, such as ability, based on rate of learning. In other words, instead of clustering growth curves we would try to cluster its derivative function. This helps better distinguish the fast learners from slow learner and from average learners in general. Also one can think of further sub-groups to explore into (such as super-smart or extra ordinary, more than average etc). This feature makes the students from different grades comparable as a smart student in grade 2 and another one in grade 8 will have different ability growth curves but may have similar learning rate curve. Finally the reason why such a classification matters is that each group may require different approach when it comes to imparting education to them. Clearly while fast learner can smoothly go over a piece of content while other two groups may be struggling to do the same content and may require more personalized treatments. So it is evident that clustering of ability growth curves based on learning rates can provide us with very useful insights.

As mentioned before learning rate is equivalent to looking at derivative function of the

growth curve but derivative may not exist always. That presents us the first challenge. Such a clustering is usually called shape-based clustering, which is dealt within the very general purview of functional clustering. In this Chapter we would propose methods suited for the purpose, that are tested to perform well on simulated data while addressing the usual challenges, that any clustering technique faces, for example number of clusters possible, label switching etc.

5.3.2 Clustering Methods

Clustering approaches are broadly of two types, a) distance based and b) model based. First type of clustering, as the name suggests, does not assume any specific model on data generation process and usually divides the data into few distinct groups by defining dissimilarity matrix or distance function. This is what is called hard clustering as every object belongs to exactly one group. On the other hand, model based clustering techniques usually assume the data to be coming from a finite mixture model. It usually computes the conditional probability of group membership and estimates of the parameters for the data generation process. Both of these approaches are widely used and the pros and cons of the approaches have been investigated in depth [Everitt (1981)]. It has also been pointed out that most of the clustering methods are suitable for use on data vectors with exchangeable, independent elements and may not be immediately applicable to longitudinal data or function data where components are repeated measurements or may share some common features [Everitt, Landau, Leese, and Stahl (2011)]. Since

here our end goal is to group students in terms of learning rate, we propose first an easy-to-implement distance-based approach and then propose a model-based approach.

5.3.3 Preview

In section 4, we elaborate and discuss various distance-based clustering methods and their limitations. Section 5 deals with common extensions of this methods to functional data or longitudinal data and role of shape and intercept in successful implementations. In section 6, we suggest spline derivative based method and discuss its advantages in IRT contexts. Section 7 discusses implementation of suggested approach in some simulated data and how the implementation differs from similar approaches in the literature. The same section eventually studies its performance as a clustering method. In section 8, a model-based alternative is suggested. Section 9 concludes with a discussion.

5.4 Distance-based Clustering Methods

Any clustering method of this kind begins by defining a similarity or dissimilarity function or distance functions between the objects it is trying to cluster. A dissimilarity or distance measure has the same properties as that of a metric except for that it may not always satisfy triangle inequality. Any metric can act as a distance measure such as, Euclidean distance or Minkowski distance etc. An example of a distance that does not

satisfy the triangle inequality is the following distance based on correlation.

$$d(x, y) = 1 - Corr(x, y).$$

Note that one of the advantages of this distance is it remains unaffected with respect to constant scale and mean shift of the vectors. Next feature of a distance-based clustering method is the algorithm it applies. There are two types of algorithms that are prevalent, *Hierarchical* and *Partition based*. Hierarchical clustering builds a tree-like structure, called dendogram, by recursively either splitting a group into smaller groups or by merging smaller groups into bigger groups. Then it decides the optimal clustering based on some properties of the tree. In the process it also defines various types of distances between the clusters (single linkage, double linkage etc). This sort of clustering is appropriate if there is hierarchy present within data vectors. In contrast partition based methods tend to map objects into K many disjoint clusters (≥ 2) by maximizing a criterion. Two popular methods include *K-means*(Hartigan and Wong (1979)) *and partitioning around medoids (PAM)* (Kaufman and Rousseeuw (2008)).

5.4.1 K-means and PAM

K-means is an unsupervised learning algorithm. Given a set of vectors $y_1, \cdots, y_n$ where $y_i \in \mathcal{R}$ for all $i = 1, \cdots, n$ K-means clustering algorithm partitions the n vectors into K sets, say, $\{ C_1, \cdots, C_K \}$ so as to minimize the sum of squared Euclidean distance to

the assigned cluster centroids denoted as

$$min \sum_{k=1}^{K} \sum_{y_i \in C_K} ||y_i - \mu_k||_2^2.$$

here $\mu_1, \cdots, \mu_K$ are the cluster centroids. To find the sets that minimize the criterion, the algorithm chooses K starting points as cluster centroids judiciously or randomly. The general K-means algorithm moves forward by alternating between two steps:

Assignment Step:

Assign each data point to its closest cluster. The kth set $C_k = \{y_i \in C_k$ if $||y_i - \mu_k||_2^2 \leq ||y_i - \mu_j||_2^2 \, \forall j \, \}$ such that every y_i is in one and only one set.

Update step:

Calculate the centroids of newly formed sets,

$$\mu_k = (\sum_{i: y_i \in C_K} y_i)/|C_K|, \quad |C_K| = \text{size of } C_k.$$

The algorithm continues to iterate until the sets no longer change. Depending on the initial partition, the algorithm is expected to converge to local optima; therefore, iterating over couple multiple random starting points can lead to a global optimum. The K-means algorithm works best when data clusters are about equal in size and shape. The algorithm seeks to find spherical clusters since the K-means algorithm is based on squared Euclidean distance. If the groups are not spherically distributed with many

spurious outliers, the computed centroids may not be representative of the clusters.

PAM algorithm attempts to improve upon some of the issues that K-means faces. The algorithm generalizes the dissimilarity matrix to user-provided any distance matrix. It is robust to outliers since the medoid or middle vector for each group is selected from the observed data vectors rather than based on mean calculations(centroids). To find K sets of vectors to minimize the sum of dissimilarity of the vectors with their respective medoids, K data points are first randomly chosen as medoids denoted as $\nu_1, \cdots, \nu_k$. The PAM algorithm alternates between the following two steps:

Build Step

Map each data point to its closest medoid based on the user-provided dissimilarity index.

Swap Step:

For each $k = 1 \cdots K$, swap the medoid ν_k, with each non-medoid observation and compute the sum of the dissimilarities of the vectors with their closest medoid. Find the configuration with the smallest sum of dissimilarities. Although it takes longer than K-means, this building and swapping procedure can return a smaller sum of dissimilarity in contrast to what the K-means algorithm achieves.

Another challenge in these clustering methods (K-means or PAM) is that one needs to specify the number of clusters in the beginning as opposed to Hierarchical clustering. In practical scenarios, this tuning parameters K, the number of clusters ($2 \leq K < n$) can be obtained by optimizing over between group and within group dissimilarities (Milligan and Cooper (1985)) or the average silhouette (Kaufman and Rousseeuw (2008)). Here

we shall discuss the average silhouette as it has broad applications to various clustering methods. For each data vector i, the silhouette $s(i)$ is defined as follows

$$s(i) = (b(i) - a(i))/max\{a(i), b(i)\}.$$

where $a(i) = d(i, C_{j_i})$, average distance between y_i and other elements of cluster C_{j_i} [C_{j_i} is the cluster y_i is assigned to]. $b(i) = min_{l:l\neq j_i} d(i, C_l)$ Define $\bar{s}$ as average of all $s(i)$'s. The chosen clustering algorithm is run and the overall average silhouette ($\bar{s}$) is calculated for each possible value of K. Then the optimal K is chosen to maximize the average silhouette while minimizing the within-group dissimilarities in comparison to the between-group dissimilarities.

5.5 Distance-based Clustering for Functional Data

5.5.1 Issues of Level and Shape

Clustering techniques that treat functional data or longitudinal data as multivariate data may fail to cluster them based on their pattern. In a study by McCoy (2010) (see also Heggeseth (2013)) people's drinking habits were studied. Two types of drinkers were chosen, heavy drinkers and beginners; for both of these, two kinds of behaviors are possible, one is increasing drinking intake over time and other is decreasing drinking intake over time. This can be observed in the following graph. When K-means is applied

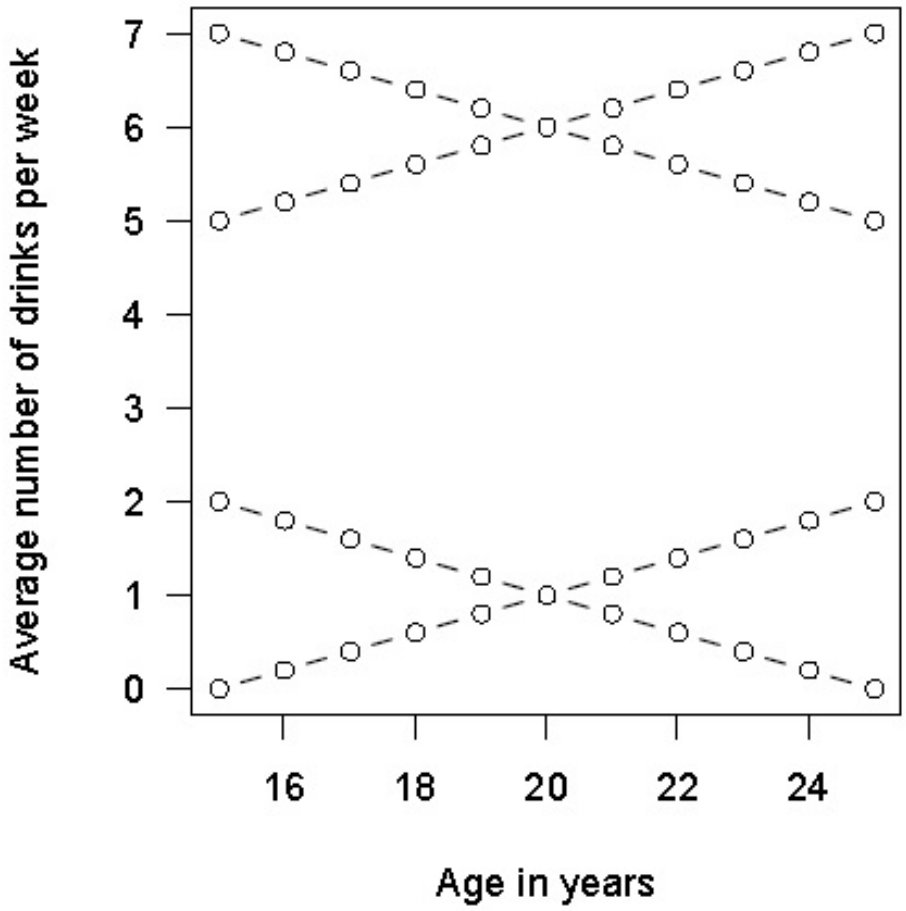


Figure 5.13: Graph of linear trajectories representing hypothetical alcohol consumption

it generates two clusters corresponding to heavy drinkers and beginners respectively. In other words, it completely ignored the overall pattern and mostly clustered based on intercept value. This seems to be general problem (as highlighted in Heggeseth (2013)) that such clustering usually ignores the trend of the curves and tends to cluster curves with similar intercept values. In the example above cohort of people were studied. So the drinking pattern can be represented as vectors of same length. Euclidean distance was considered for K-means. To particularly capture shape two types of approaches were mostly popular that ignore the intercept effects. 1> Derivative-based approaches, which we eventually propose in this Chapter and 2> using a distance that is invariant to level changes (like correlation-based)

5.5.2 Derivative-based Approaches

There are many approaches as one can estimate derivative in many ways and use that to cluster. Given data points one can compute consecutive difference quotients to use them as derivative estimates and define distances on these estimates. But this may lead to large variance (D'Urso (2000)). A more popular approach is to project the data to"good " class of functions which are differentiable and use the derivative of the projected functions. Tarpey (2012) used class of Fourier transforms for differentiable class while Zerbe (1979) worked with polynomial class. Both fourier transforms and polynomial classes are restrictive (as they are both infinitely differentiable). A larger class of differentiable functions is class of splines, Heggeseth (2013) extended the idea to

splines. Our methodology is closely based on Heggeseth (2013).

5.5.3 Extension to Longitudinal Data of Various Lengths

Let us assume that we have longitudinal data points, $y_{i,j}$, where $i = 1, \cdots, n$ and $j = 1, \cdots, T_i$. We further assume that there is an underlying function or process $f_i(t)$ for all $i = 1, \cdots, n$ and $y_{i,j}$'s are repeated measures collected from this process at certain designated time-points. If T_i's are the same one can apply K-means or PAM based on Euclidean distances between the vectors. Alternatively if the functional forms are known on can apply K-means based on L^2 distance between two functions like the following.

$$||L^2(f_1, f_2)|| = [\int_{\mathcal{T}} (f_1(t) - f_2(t))^2 dt,]^{1/2}, \quad \mathcal{T} \text{ :period of measurements.} \tag{5.1}$$

Note T_i's are not the same. So the first method falls apart. But if data is projected to functional space 2nd method is still applicable. This explains why D'Urso (2000) method is less popular although it might give unbiased estimates of derivatives.

5.6 Clustering Shapes Based on Derivatives of Spline Estimates

Let $y_{i,j}$'s be defined as in section 5.5.3. We intend to cluster underlying processes, $f_i(t)$'s based on their shapes. In this case we define shape by rates of changes of values and we

outline a method very similar to Heggeseth (2013). The algorithm is given below.

1: **procedure** : DERIVATIVE BASED CLUSTERING

2: Define B-spline basis of order 3 with 2 equi-distant internal knots splines on the measurement period. (Review 4.1.2 for splines). Let them be $B_{j,3}(t), \quad j = 1, \cdots, 5$.

3: Obtain Least Squre estimate of $f_i(t), \quad \hat{f_i(t)} = \sum_j \alpha_{i,j} B_{j,3}(t) \quad i = 1, \cdots, n$.

4: Obtain derivative spline coefficients (Pleas see section 4.3 for exact expression). Lets call them $\alpha^*_{i,j} \quad j = 1, \cdots, 4$. Define $\boldsymbol{\alpha}^*_i = (\alpha^*_{i,1}, \cdots, \alpha^*_{i,4})'$.

5: Apply K-means based on the Euclidean distance between $\boldsymbol{\alpha}^*_i$'s.

6: Choose the number of clusters based on silhouette method.

Note that this method of working with L^2 metric between derivative spline function is equivalent to L^2 metric between some linear combinations of derivative spline coefficient.

$$
\begin{aligned}
L^2(f_1', f_2') &= \int_{\mathcal{T}} (f_1'(t) - f_2'(t))^2 dt, , \quad \mathcal{T} \text{ :period of measurements} \\
&= \int_{\mathcal{T}} [\sum_j (\alpha^*_{1,j} - \alpha^*_{2,j}) B_{j,3}(t)]^2 dt, \\
&= \int_{\mathcal{T}} [\sum_{j_1, j_2} [(\alpha^*_{1,j_1} - \alpha^*_{2,j_1}) B_{j_1,3}(t)][(\alpha^*_{1,j_2} - \alpha^*_{2,j_2}) B_{j_2,3}](t) dt, \\
&= \sum_{j_1, j_2} (\alpha^*_{1,j_1} - \alpha^*_{2,j_1})(\alpha^*_{1,j_2} - \alpha^*_{2,j_2}) \int_{\mathcal{T}} B_{j_1,3}(t) B_{j_2,3}(t) dt, \\
&= (\boldsymbol{\alpha}^*_1 - \boldsymbol{\alpha}^*_2)' \Sigma (\boldsymbol{\alpha}^*_1 - \boldsymbol{\alpha}^*_2), \quad \Sigma_{j_1, j_2} = \int_{\mathcal{T}} B_{j_1,3}(t) B_{j_2,3}(t) dt, \\
&\quad \Sigma \text{ is positive semi-definite matrix here.}
\end{aligned}
$$

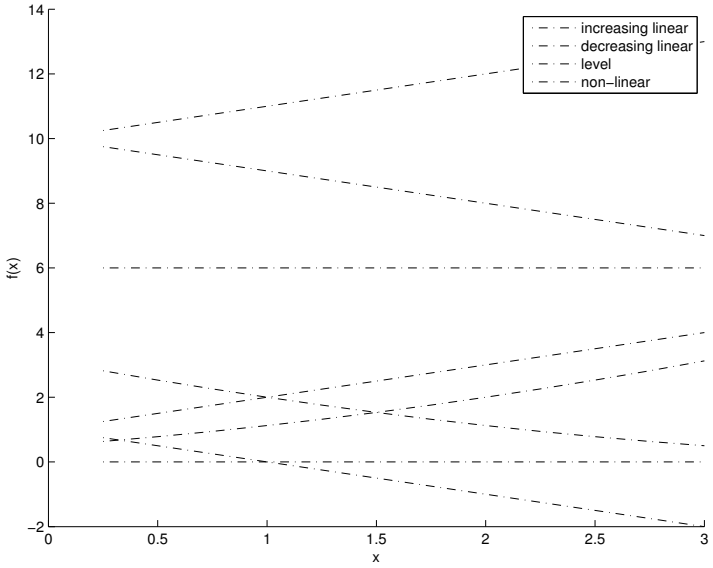


Figure 5.14: Mean functions

5.7 Performance of the Proposed Method in Simulation Study

We consider 8 different mean functions on [0,3] as follows:

$$f_1(t) = 10 + t; \quad f_2(t) = 10 - t; \quad f_3(t) = 1 + t; \quad f_4(t) = 1 - t$$
$$f_5(t) = 0; \quad f_6(t) = 6, \quad f_7(t) = \frac{(5-t)^2}{8}, \qquad f_8(t) = \frac{(2+t)^2}{8}.$$

Sample curves were generated with various levels of Gaussian noise (σ level) . Both mean functions and spline estimates of sample curves were plotted for $\sigma = 0.25$ and $\sigma = 0.5$ and for the case of one inner knot (See Figure 5.14 and Figure 5.15) .

In terms of shape there are mostly three types of curves; increasing , decreasing

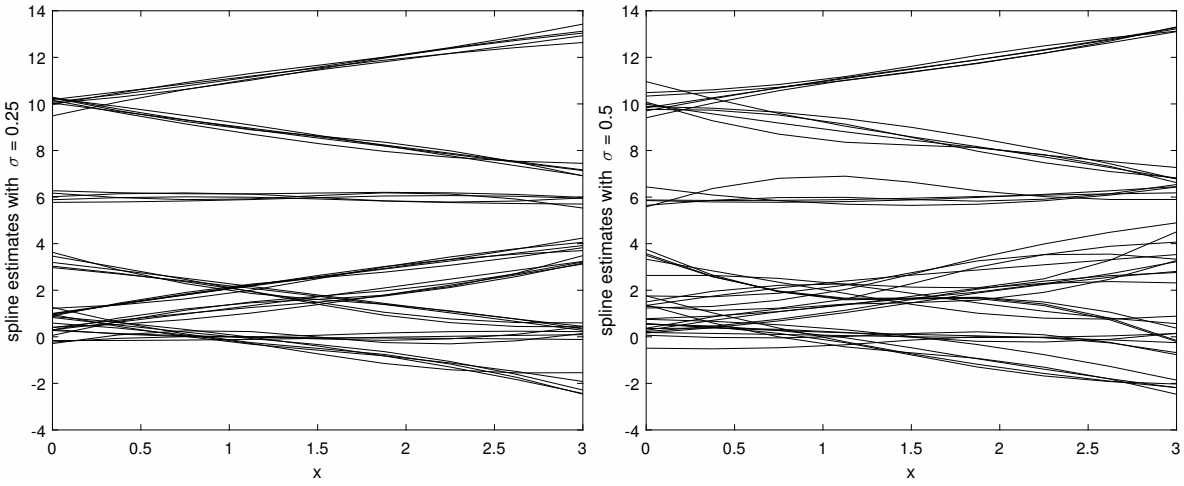


Figure 5.15: Spline estimates, $\sigma = 0.025$ (left), $\sigma = 0.5$(right)

and level, being considered above. Algorithm given in 1 is applied on the data and mis-classification rates are computed to understand goodness of the performance. Since assigned labels may be mis-matched, cluster labels are validated through permuting within groups until maximum number of matches between labels is obtained. This is maximization of trace of $3\times$ 3 match matrix under all permutations of labels. K-means was implemented with 5 random sets of starting values. This is very similar method as suggested by Heggeseth (2013) except for the fact that we considered non-linear function in our simulation study and B-splines were computed based on equi-spaced internal knots as opposed to quartile-based knots. Also Heggeseth (2013) did not consider any internal knots. This simulation is performed for various noise level and with two different sequences of equi-spaced internal knots. Performance results are summarized below for each of these cases through misclassification rates. Clearly from the table it is evident that mis-classification increases as noise level and number of internal knots increase as

σ	Inner knots	Misclassification Rate
0 (True Curves)	1	0%
0.1	1	0%
0.25	1	7.5%
0.5	1	32.5%
0 (True Curves)	2	0%
0.1	2	0%
0.25	2	30%
0.5	2	35%

Table 5.8: Misclassification Increases with Noise and Number of Knots

with more internal knots, splines tend to over-fit the noise.

5.8 Model-based Alternative

Alternatively one can incorporate the cluster structure within the modeling framework in couple of ways. The easiest way is to consider c_i (as in the DIR-RT 2nd stage model) as the representative for average growth rate and hence different learning patterns would correspond to different values of c_i's. If we assume there are 3 clusters the we can put a prior on c_i as a mixture of Gaussians.

$$c_i \sim \sum_{k=1}^{3} \pi_k f(\mu_k, \sigma_k), \;\; f : \text{Gaussian}.$$

We then calculate the posterior probability that c_i belongs to a particular group. There will be label switching or identifiability issues, which can be mitigated by putting appropriate constraints. Also if the number of clusters are not known we can vary the number

of clusters and consider them as competing models and choose the one based on some model selection criterion.

5.9 Applications to MetaMetrics Test Data

Currently, both the methodologies are being implemented to MetaMetrics test data. The distance based approach is applied to cluster growth curve estimates of DIR-RT models, as obtained by posterior median. After that cluster results are mapped with c_i values. We choose K to be 3; this is based on the belief that there are 3 types of students; fast learners, slow learners and the average. From eye-balling it is clear that high growth curves and low ones are well separated. Alternatively, we could have used silhouette method to decide from a set of reasonable values. On the other hand for model-based approach the modeling framework of 2nd stage of DIR-RT is modified. Along with that full conditionals are being modified.

5.10 Discussion

We verified the goodness of the suggested distance-based approach in light of simulated data. In the simulation study we dealt with, both true curves (σ =0) and samples of true curves with various levels of added noise. We note that with the increase of noise level performance of clustering technique becomes poor as splines tend to over-fit. Currently the same is applied to DIR-RT estimates of latent ability, thereby clustering the latent

curves. We believe that the cluster should reflect diversity represented by estimates of average growth rates, $c_i's$. A derivative Spline based clustering can help differentiate between curves based on curvature and help identify different groups of students. The distance-based approach is fast and easy-to-implement where model based approach would require MCMC steps. Hence the latter is going to take longer time to implement.

Appendix A

MCMC Computation for DIR-RT Models

MCMC proceeds by running through block Gibbs sampler, thereby updating a block of variables at a time. The following lists the steps for sampling block of variables from their full conditional posterior distribution.

Step 1: **Sampling Y: Truncated Normal Distribution Sampling**

Given θ, φ, η, γ, and X the latent variables $\{Y_{i,t,s,l}\}$ are sampled from

$$Y_{i,t,s,l} \sim \mathcal{N}_{+}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 1$$

$$Y_{i,t,s,l} \sim \mathcal{N}_{-}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 0\,,$$

where $\mathcal{N}_{+}(\cdot,\cdot)$ means the normal distribution truncated at the left by zero while $\mathcal{N}_{-}(\cdot,\cdot)$ is the normal distribution truncated at the right by zero.

Step 2: **Sampling θ: The Sampling Scheme Depends on the Choice of $L(\cdot)$**

Step 2.1: $L(\cdot) = \cdot$**, Forward Filtering and Backward Sampling (FFBS).**

Define $Z_{i,t,s,l} = Y_{i,t,s,l} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s} - \rho^{-1}$, $H_{i,t,s} = \log(R_{i,t,s}) - \mu_i - \nu_{i,t} + \beta(a_{i,t,s} -$

ρ^{-1}) and use the setting of $\lambda_{i,t} = \theta_{i,t} - \rho^{-1}$ and $g_{i,t} = 1 - c_i\rho\Delta^+_{i,t}$, then the (conditional) one-parameter DIR-RT model will fit the framework of dynamic linear model (see the reference, West and Harrison (1997)), i.e.,

$$\begin{aligned} \text{System Equation:} \qquad & \lambda_{i,t} = g_{i,t}\lambda_{i,t-1} + w_{i,t}, \\ \text{Observation Equation:} \qquad & Z_{i,t,s,l} = \lambda_{i,t} + \xi_{i,t,s,l}, \\ & H_{i,t,s} = \beta\lambda_{i,t} + \zeta_{i,t,s}, \end{aligned}$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\xi_{i,t,s,l} \sim \mathcal{N}(0, 4\gamma^2_{i,t,s,l} + \sigma^2)$ and $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$. Define information available on the tth day as

$$\mathcal{F}_{i,t} = \left\{ g_{i,e}, \phi, \psi, \varphi, \eta, \varrho, c, \beta, \mu_i, \nu_{i,e}, H_{i,e,1}, \cdots, H_{i,e,S_{i,e}}, Z_{i,e,1,1}, \cdots, Z_{i,e,S_{i,e},K_{i,e,S_{i,e}}} \right\}^t_{e=1}.$$

The FFBS algorithm can be implemented to block update each $\lambda_i = (\lambda_{i,0}, \cdots, \lambda_{i,T_i})'$:

1. (Forward Filtering) For $t \geq 1$, it is not hard to show that $[\lambda_{i,t} \mid \mathcal{F}_{i,t}] \sim \mathcal{N}(\mu_{i,t}, V_{i,t})$, with $\mu_{i,t} = V_{i,t}(R^{-1}_{i,t}d_{i,t} + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}} Z_{i,t,s,l}\psi_{i,t,s,l} + \varrho\beta\sum_{s=1}^{S_{i,t}} H_{i,t,s})$ and $V_{i,t} = \left(R^{-1}_{i,t} + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}}\psi_{i,t,s,l} + \varrho\beta^2 S_{i,t}\right)^{-1}$. Notice when $t = 0$, $\lambda_{i,0}$ follows $\mathcal{N}(\mu_{i,0}, V_{i,0})$ with $\mu_{i,0} = \mu_{G_j} - \rho^{-1}$ and $V_{i,0} = V_{G_j}$. Here, from system equation it follows the prior distribution of $\lambda_{i,t} \mid \mathcal{F}_{i,t-1} \sim \mathcal{N}(d_{i,t}, R_{i,t})$, where $d_{i,t} = g_{i,t}\mu_{i,t-1}$ and $R_{i,t} = g^2_{i,t}V_{i,t-1} + \phi^{-1}\Delta_{i,t}$.

2. (Backward Sampling) Save all quantities of $\mu_{i,t}$ and $V_{i,t}$. Then, draw λ_{i,T_i} from

$\mathcal{N}(\mu_{i,T_i}, V_{i,T_i})$. When $t = (T_i - 1)$ to 0, with some algebra, we can see $\lambda_{i,t}$ will be drawn from $[\lambda_{i,t} \mid \lambda_{i,t+1}, \mathcal{F}_{i,t}] \sim \mathcal{N}(h_{i,t}, m_{i,t})$, where $h_{i,t} = m_{i,t}(V_{it,}^{-1}\mu_{i,t} + \phi g_{i,t+1}\Delta_{i,t+1}^{-1}\lambda_{i,t+1}$) and $m_{i,t} = (\phi g_{i,t+1}^2\Delta_{i,t+1}^{-1} + V_{i,t}^{-1})^{-1}$.

Thus, for $t = 0, \cdots, T_i$, set $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$ and θ_i is sampled as a whole block, noticing $\Pr(\theta_i \mid \mathcal{F}_{i,T_i}) = \Pr(\theta_{i,T_i} \mid \mathcal{F}_{i,T_i})\Pr(\theta_{i,T_i-1} \mid \theta_{i,T_i}, \mathcal{F}_{i,T-1}) \cdots \Pr(\theta_{i,0} \mid \theta_{i,1}, \mathcal{F}_{i,0})$.

Step 2.2 $L(\cdot) = |\cdot|$**, Conditional Mixture of Truncated Normal Distribution**

Consider ϕ, c, Y, φ, η, γ, μ, ν, ϱ and β are given. Similarly, the (conditional) one-parameter DIR-RT model fits the framework of state space models,

$$\begin{aligned} \text{System Equation:} \qquad & \lambda_{i,t} = g_{i,t}\lambda_{i,t-1} + w_{i,t}, \\ \text{Observation Equation:} \qquad & Z_{i,t,s,l} = \lambda_{i,t} + \xi_{i,t,s,l}, \\ & H_{i,t,s} = \beta|\lambda_{i,t} - q_{i,t,s}| + \zeta_{i,t,s}, \end{aligned}$$

where $q_{i,t,s} = a_{i,t,s} - \rho^{-1}$ and other parameters have same definitions as before. Instead of sampling $\theta_{i,t}$, we are going to sample $\lambda_{i,t}$ and then $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$. A Gibbs algorithm to sample $\lambda_{i,0}, \ldots, \lambda_{i,T_i}$ is designed below.

For $t = 1, \cdots, T_i - 1$, after some mathematical derivations, $\lambda_{i,t}^{(M)}$ can be regarded as drawing from a mixture of truncated normal distribution, i.e.,

$$\lambda_{i,t}^{(M)} \sim \Pr(\lambda_{i,t} \mid \lambda_{i,t-1}^{(M)}, \lambda_{i,t+1}^{(M-1)}, \mathcal{F}_{i,T_i}),$$

where $\Pr(\lambda_{i,t} \mid \lambda_{i,t-1}^{(M)}, \lambda_{i,t}^{(M-1)}, \mathcal{F}_{i,T_i}) = \sum_{s=0}^{S_{i,t}} p_{i,t,s}\mathcal{N}_{(q_{i,t,s},q_{i,t,s+1}]}(d_{i,t,s}, R_{i,t})$ with $q_{i,t,0} = -\infty$, $q_{i,t,S_{i,t}+1} = \infty$, $p_{i,t,s}$ defined as $p_{i,t,s} = \frac{\Phi\left(\frac{q_{i,t,s}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)-\Phi\left(\frac{q_{i,t,s+1}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)}{\sum_{s=0}^{S_{i,t}}\left(\Phi\left(\frac{q_{i,t,s}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)-\Phi\left(\frac{q_{i,t,s+1}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)\right)}$, and $m_{i,t} = \phi\Delta_{i,t}^{-1}g_{i,t}\lambda_{i,t-1}^{(M)} + \phi\Delta_{i,t+1}^{-1}g_{i,t+1}\lambda_{i,t+1}^{(M-1)} + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}}\psi_{i,t,s,l}Z_{i,t,s,l} + \varrho\sum_{s=1}^{S_{i,t}}\beta^2 q_{i,t,s}$, $d_{i,t,s} = R_{i,t}\left(m_{i,t} + \varrho\beta\left(\sum_{j=0}^{s}H_{i,t,j} - \sum_{j=s}^{S_{i,t}}H_{i,t,j}\right)\right)$, $s = 0, \cdots, S_{i,t}$, $H_{i,t,0} = 0$, $R_{i,t} = \left(\phi\Delta_{i,t}^{-1} + \phi\Delta_{i,t+1}^{-1}g_{i,t+1}^2 + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}}\psi_{i,t,s,l} + \varrho\beta^2 S_{i,t}\right)^{-1}$. The formula is almost the same for sampling λ_{i,T_i} with only deleting the terms involving the index of $t+1$ in $m_{i,t}$ and $R_{i,t}$ and similarly, deleting the terms involving the index of $t-1$ in $m_{i,t}$ and $R_{i,t}$ for $\lambda_{i,0}$. At the end, set $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$, for $t = 1, \cdots, T_i$.

Step 3: **Sampling c: Truncated Normal Distribution Sampling**

When θ and ϕ are given, the full conditional distribution of c_i is the truncated normal distribution

$$c_i \sim \mathcal{N}_+\left(\frac{\sum_{t=1}^{T_i}(1-\rho\theta_{i,t-1})(\theta_{i,t}-\theta_{i,t-1})\Delta_{i,t}^{+}\Delta_{i,t}^{-1}}{\sum_{t=1}^{T_i}(\Delta_{i,t}^{+}(1-\rho\theta_{i,t-1}))^2\Delta_{i,t}^{-1}}, \frac{1}{\phi\sum_{t=1}^{T_i}(\Delta_{i,t}^{+}(1-\rho\theta_{i,t-1}))^2\Delta_{i,t}^{-1}}\right).$$

Step 4: **Sampling η: Multivariate Normal Distribution Sampling**

When θ, φ, τ, Y and γ are given, if $S_{i,t} = 1$, $\eta_{i,t,S_{i,t}} = 0$, while if $S_{it} > 1$, then the full conditional distribution of η_{it}^* is the multivariate normal distribution

$$\eta_{i,t}^* \sim \mathcal{N}_{S_{i,t}-1}\left((A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}Y_{i,t}^*, (A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}\right),$$

where $Y_{i,t}^* = (Y_{i,t,1,1} - \theta_{i,t} + a_{i,t,1} - \varphi_{i,t}, \cdots, Y_{i,t,1,K_{i,t,1}} - \theta_{i,t} + a_{i,t,K_{i,t,1}} - \varphi_{i,t}, \cdots, Y_{i,t,S_{i,t},K_{i,t,S_{i,t}}} -$

$\theta_{i,t} + a_{i,t,K_{i,t,S_{i,t}}} - \varphi_{i,t})'$, $\Sigma_{\psi_{i,t}}^{-1} = \text{diag}((\psi_{i,t,1,1}, \cdots, \psi_{i,t,S_{i,t},K_{i,t,S_{i,t}}})')$, $A_{i,t} = \left(\bigoplus_{s=1}^{S_{i,t}-1} \mathbf{1}'_{K_{i,t,s}}, -J_{K_{i,t,S_{i,t}} \times (S_{i,t}-1)}\right)'$ with $\mathbf{1}_{K_{i,t,s}}$ being an unit vector with $K_{i,t,s}$ dimension, $\bigoplus$ indicating direct sum. Set $\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s}$.

Step 5: **Sampling τ: Gamma Distribution Sampling**

When η is given, the full conditional distribution of τ_i is the gamma distribution

$$\tau_i \sim \mathcal{G}a\left(\frac{\sum_{t=1}^{T_i} S_{i,t} - (T_i + 1)}{2}, \frac{\sum_{t=1}^{T_i} \eta_{i,t}^{*\prime} \Sigma_{i,t}^{-1} \eta_{i,t}^{*}}{2}\right).$$

where $\mathcal{G}a(a, b)$ denotes a gamma distribution with the shape parameter a and the rate parameter b.

Step 6: **Sampling φ: Normal Distribution Sampling** When θ, η, Y and γ are given, the full conditional distribution of φ_{it} is the normal distribution

$$\varphi_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \eta_{i,t,s})}{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} + \delta_i}, \frac{1}{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{its}} \psi_{i,t,s,l} + \delta_i}\right).$$

Step 7: **Sampling δ: Gamma Distribution Sampling** When φ is given, the full conditional distribution of δ_i is the gamma distribution

$$\delta_i \sim \mathcal{G}a\left(\frac{T_i - 1}{2}, \frac{\sum_{t=1}^{T_i} \varphi_{i,t}^2}{2}\right).$$

Step 8: **Sampling ϕ: Gamma Distribution Sampling**

When θ, c is given, the full conditional distribution of ϕ is the gamma distribution

$$\phi \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n} T_i - 1}{2}, \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} \Delta_{i,t}^{-1}(\theta_{i,t} - \theta_{i,t-1} - c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+})^2}{2}\right).$$

Step 9: **Sampling γ : Metropolis-Hastings Sampling**

Given Y, θ, φ and η, the full conditional distribution of $\gamma_{i,t,s,l}$ is not in a closed form. Thus, we resort to a Metropolis-Hastings scheme to sample this distribution. A suitable proposal for sample γ is K-S distribution itself. Thus, we first sample γ from the K-S distribution and then let

$$\gamma_{i,t,s,l}^{(M)} = \begin{cases} \gamma^*, & \text{with probability} \quad \min(1, LR) \\ \gamma_{i,t,s,l}^{(M-1)}, & \text{otherwise} \end{cases}$$

where, given Y, θ, φ and η,

$$\begin{aligned} LR &= \sqrt{\frac{\sigma^2 + 4(\gamma_{i,t,s,l}^{(M-1)})^2}{\sigma^2 + 4(\gamma^*)^2}} \exp\left\{-\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2}\right. \\ &\cdot \left.\left(\frac{1}{\sigma^2 + 4(\gamma^*)^2} - \frac{1}{\sigma^2 + 4(\gamma_{itsl}^{(M-1)})^2}\right)\right\}. \end{aligned}$$

Step 10: **Sampling μ: Normal Distribution**

Given ϱ, ν, θ, β, the full conditional distribution of μ_i is the normal distribution,

$$\mu_i \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T_i}\sum_{s=1}^{S_{i,t}} [\log(R_{i,t,s}) + \nu_{i,t} - \beta L(\theta_{i,t} - a_{i,t,s})]}{\sum_{t=1}^{T_i} S_{i,t}}, \frac{1}{\varrho\sum_{t=1}^{T_i} S_{i,t}}\right).$$

Step 11: **Sampling ν: Normal Distribution**

Given ϱ, μ, θ, β, the full conditional distribution of $\nu_{i,t}$ is the normal distribution,

$$\nu_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}} -\varrho\left[\log(R_{i,t,s}) - \mu_i - \beta L(\theta_{i,t} - a_{i,t,s})\right]}{\varrho S_{i,t} + \kappa_i}, \frac{1}{\varrho S_{i,t} + \kappa_i}\right).$$

Step 12: **Sampling ϱ: Gamma Distribution**

Given μ, ν, θ, β, the full conditional distribution of ϱ is the gamma distribution,

$$\varrho \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} S_{i,t}}{2}, \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{s=1}^{S_{i,t}}\left[\log(R_{i,t,s}) - \mu_i + \nu_{i,t} - \beta L(\theta_{i,t} - a_{i,t,s})\right]^2}{2}\right).$$

Step 13: **Sampling κ: Gamma Distribution**

Given ν, the full conditional distribution of κ_i is the gamma distribution,

$$\kappa_i \sim \mathcal{G}a\left(\frac{T_i}{2}, \frac{\sum_{t=1}^{T_i}\nu_{i,t}^2}{2}\right).$$

Step 15: **Sampling β: Normal Distribution**

Given ϱ, ν, μ, and θ the full conditional distribution of β is

$$\beta \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{s=1}^{S_{i,t}} L(\theta_{i,t} - a_{i,t,s})\left[\log(R_{i,t,s}) - \mu_i + \nu_{i,t}\right]}{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{s=1}^{S_{i,t}}(\theta_{i,t} - a_{i,t,s})^2}, \frac{1}{\varrho\sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{s=1}^{S_{i,t}}(\theta_{i,t} - a_{i,t,s})^2}\right).$$

Appendix B

DIC Computation based on Partial DIR-RT Models

Let us define the vectorized parameters $\Theta = vec(\varrho, \kappa, \beta, \theta)$. Set $R^{\star}_{i,t} = (\log R_{i,t,1} \cdots , \log R_{i,t,S_{i,t}})'$ and $\mu_{i,t} = (\mu_i + \beta L(\theta_{i,t} - a_{i,t,1}), \cdots , \mu_i + \beta L(\theta_{i,t} - a_{i,t,S_{i,t}}))'$. We can show that $R^{\star}_{i,t} \overset{ind}{\sim} \mathcal{N}_{S_{it}}(\mu_{i,t}, \Omega_{i,t})$, where ind indicates independent and $\Omega_{i,t} = \kappa_i^{-1} J_{S_{i,t}} + \varrho^{-1} I_{S_{i,t}}$. The partial likelihood of DIR-RT models based only on response times is then

$$L(\mathbf{\Theta}|R^{\star}_{i,t}) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \phi(R^{\star}_{i,t}|\Theta),$$

where $\phi(\cdot|\Theta)$ is the multivariate normal probability density function. Then, according to (3.2), the partial DIC defined for the goodness of fit and model complexity of the part of response times in DIR-RT models is

$$DIC_P = 2E_{\Theta|R^{\star}_{i,t}} Q(\Theta, R^{\star}_{i,t}, L(\cdot)) - Q(\bar{\Theta}, R^{\star}_{i,t}, L(\cdot)),$$

with DIC_P implying partial DIC and $Q(\Theta, R^{\star}_{i,t}, L(.)) = -2logL(\mathbf{\Theta}|R^{\star}_{i,t})$. Immediately following from the facts of $|\Omega_{i,t}| = \varrho^{-S_{i,t}}(1 + \varrho\frac{S_{i,t}}{k_i})$, $\Omega^{-1}_{i,t} = \varrho I_{S_{i,t}} - \frac{\varrho^2}{k_i+\varrho S_{i,t}} J_{S_{i,t}}$, we can simplify

$$\begin{aligned}
-2logL(\mathbf{\Theta}|R^{\star}_{i,t}) &= \sum_i \sum_t [(R^{\star}_{i,t} - \mu_{i,t})'\Omega^{-1}_{i,t}(R^{\star}_{i,t} - \mu_{i,t}) + log|\Omega_{i,t}| + (S_{i,t}log2\pi)] \\
&= \varrho \sum_{i,t} (R^{\star}_{i,t} - \mu_{i,t})'(R^{\star}_{i,t} - \mu_{i,t}) - \varrho^2 \sum_{i,t} \frac{[(R^{\star}_{i,t} - \mu_{i,t})'\mathbf{1}_{S_{i,t}}]^2}{k_i + \varrho S_{i,t}} \\
&+ \log(\frac{2\pi}{\varrho})(\sum_{i,t} + S_{i,t}) \sum_{it} \log(\kappa_i + \varrho S_{i,t}) - \sum_{i,t} \log \kappa_i.
\end{aligned}$$

Appendix C

MCMC Computations for DIR-SMSG Models

Full conditionals when ϕ is unknown

Step 1:**Sampling Y: Truncated Normal Distribution Sampling**

$$Y_{i,t,s,l} \sim \mathcal{N}_{+}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 1$$

$$Y_{i,t,s,l} \sim \mathcal{N}_{-}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 0\,,$$

where $\mathcal{N}_{+}$ means the normal distribution truncated at the left by zero while $\mathcal{N}_{-}$ is the normal distribution truncated at the right by zero and $\psi^{-1}_{i,t,s,l} = 4\nu^2_{i,t,s,l} + \sigma^2$. Sampling from truncated normals is fast and easy.

Step 2:**Sampling $\boldsymbol{\theta}$: normal sampling and computation of Z**

Define: $Z_{i,t,s,l} = Y_{i,t,s} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s}$. $\quad Z_{i,t,.,.} = \sum_{s,l} Z_{i,t,s,l}$. $\quad (\psi Z)_{i,t,s,l} = \psi_{i,t,s,l} Z_{i,t,s,l}$.

$$\pi(\theta_{i,t} \mid .) \;\sim\; \mathcal{N}(R_{it}(X_i(t,:)\boldsymbol{\alpha}_i\phi\Delta_{i,t}^{-1} + (\psi Z)_{i,t,.,.}), R_{it} = [\phi/\Delta_{i,t} + \psi_{i,t,.,.}]^{-1}) \quad \text{(C.1)}$$

Step 3:**Sampling $\boldsymbol{\alpha}$: Truncated multivariate normal sampling**

$$\begin{aligned} \pi(\boldsymbol{\alpha}_i \mid .) &\propto \exp^{-(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^m)' P_i (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^m)/2} \prod_{j\geq 2}^{m} \mathbf{1}(\alpha_{i,j} \geq \alpha_{i,j-1}) \qquad \text{(C.2)} \\ P_i &= \phi \sum_t [X_i(t,:)' X_i(t,:)] \Delta_{i,t}^{-1} + \omega_i K^{\delta}, \quad \boldsymbol{\alpha}_i^m = P_i^{-1} \phi \sum_t [X_i(t,:)' \theta_{i,t} \Delta_{i,t}^{-1}] \end{aligned}$$

Since pdf has a domain restriction, it leads to multivariate truncated normal. We follow Robert (1995)'s approach to run a single-move MCMC chain to sample $\boldsymbol{\alpha}_i$. Let $\boldsymbol{\alpha}_i^c$ be the current state of MCMC chain. Let us run the sub-chain upto L (taken to be 100) many steps. The algorithm can be given for an individual i as follows.

Set $\boldsymbol{\alpha}_i^{(0)} = \boldsymbol{\alpha}_i^c$. Draw samples from the univariate truncated normals successively and repeat the following m steps for l =1 through L.

$$\begin{aligned} 1.&\quad \alpha_{i,1}^{(l)} \sim \mathcal{N}(\mu_{i,1}, \sigma_{i,1}^2, -\infty, \alpha_{i,2}^{(l-1)}) \\ 2.&\quad \alpha_{i,2}^{(l)} \sim \mathcal{N}(\mu_{i,2}, \sigma_{i,2}^2, \alpha_{i,1}^{(l)}, \alpha_{i,3}^{(l-1)}) \\ &\quad \vdots \\ m.&\quad \alpha_{i,m}^{(l)} \sim \mathcal{N}(\mu_{i,m}, \sigma_{i,m}^2, \alpha_{i,m-1}^{(l)}, +\infty) \end{aligned}$$

where $\mathcal{N}(\mu, \sigma^2, \mu_l, \mu_r)$ denotes a Gaussian distribution with mean μ, variance σ^2 and with left truncation point μ_l and right truncation point μ_r, respectively. The truncation points in the algorithm above are the current states of the adjacent parameters. The parameters μ_l and σ_l^2 are defined as follows. for $j = 1, .., m$ (the index i is suppressed in the following)

$$\begin{aligned} \mu_j &= \alpha_j^m - P_{jj}^{-1}\{\sum_{j'<j}(\alpha_{j'}^{(l)} - \alpha_{j'}^m)P_{jj'} + \sum_{j'>j}(\alpha_{j'}^{(l-1)} - \alpha_{j'}^m)P_{jj'}\} \\ \sigma_j^2 &= P_{jj}^{-1}. \end{aligned}$$

are the conditional means and variances of the (non-truncated) normal posterior.

Step 4:**Sampling $\boldsymbol{\omega}$: Gamma sampling**

$$\pi(\omega_i \mid .) \sim \mathcal{G}a(a + (m-2)/2, \boldsymbol{\alpha_i}' K^\delta \boldsymbol{\alpha_i}/2 + b)$$

Step 5:**Sampling $\boldsymbol{\tau}$: Gamma sampling**

$$\pi(\tau_i \mid .) \sim \mathcal{G}a(\frac{S_{i,.} - T_i}{2} - 1/2, \sum_t \eta_{i,t}^{*\prime}\Sigma_{i,t}^{-1}\eta_{i,t}^*/2), \quad \Sigma_{i,t}^{-1} = I_{S_{i,t}-1} + J_{S_{i,t}-1}$$

Step 6:**Sampling $\boldsymbol{\eta}^*$: Multivariate Normal Distribution Sampling** If $S_{i,t} > 1$, then the full conditional distribution of $\boldsymbol{\eta}_{i,t}^*$ is the multivariate normal distribution

$$\boldsymbol{\eta}_{i,t}^* \sim \mathcal{N}_{S_{i,t}-1}\left((A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}\boldsymbol{Y}_{i,t}^*, (A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}\right),$$

where $\boldsymbol{Y}^*_{i,t} = (Y_{i,t,1,1} - \theta_{i,t} + a_{i,t,1} - \varphi_{i,t}, \cdots, Y_{i,t,1,K_{i,t,1}} - \theta_{i,t} + a_{i,t,1} - \varphi_{i,t}, \cdots, Y_{i,t,S_{i,t},K_{i,t,S_{i,t}}} - \theta_{i,t} + a_{i,t,K_{i,t,S_{i,t}}} - \varphi_{i,t})'$, $\Sigma^{-1}_{\psi_{i,t}} = \text{diag}((\psi_{i,t,1,1}, \cdots, \psi_{i,t,S_{i,t},K_{i,t,S_{i,t}}})')$,

$$A_{i,t} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix}_{(\sum_{s=1}^{S_{i,t}} K_{i,t,s}) \times (S_{i,t}-1)},$$

and $\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s}$. When $S_{i,t} = 1$, $\eta_{i,t,S_{i,t}} = 0$.

Step 7:**Sampling $\boldsymbol{\varphi}$: Normal Distribution Sampling**

$$\varphi_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \eta_{i,t,s})}{\psi_{i,t,.,.} + \delta_i}, \frac{1}{\psi_{i,t,.,.} + \delta_i} \right).$$

Step 8:**Sampling $\boldsymbol{\delta}$: Gamma Distribution Sampling**

When φ is given, the full conditional distribution of δ_i is the gamma distribution

$$\delta_i \sim \mathcal{G}a\left(\frac{T_i-1}{2}, \frac{\sum_{t=1}^{T_i}\varphi_{i,t}^2}{2}\right).$$

Step 9:**Sampling ϕ: Gamma Distribution Sampling**

$$\phi \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n}T_i-1}{2}, \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\Delta_{i,t}^{-1}(\theta_{i,t}-X_i(t,:)\boldsymbol{\alpha}_i)^2}{2}\right).$$

Step 10:**Sampling $\boldsymbol{\gamma}$: Metropolis-Hastings Sampling**

$$\pi(\gamma_{i,t,s,l}|.) \propto \sqrt{\frac{1}{\sigma^2+4\gamma_{i,t,s,l}^2}}\exp\left\{-\frac{(Y_{i,t,s,l}-\theta_{i,t}+a_{i,t,s}-\varphi_{i,t}-\eta_{i,t,s})^2}{2(\sigma^2+4\gamma_{i,t,s,l}^2)}\right\}p(\gamma_{i,t,s,l}), \tag{C.3}$$

which is not in closed form. So we shall resort to a Metropolis-Hastings scheme to sample this distribution. A suitable proposal for sample γ is K-S distribution itself. Thus, we first sample γ from the K-S distribution. Then, we let

$$\gamma_{i,t,s,l}^{(M)} = \begin{cases} \gamma^*, & \text{with probability} \quad \min(1, LR) \\ \gamma_{i,t,s,l}^{(M-1)}, & \text{otherwise} \end{cases}$$

where,

$$LR = \sqrt{\frac{\sigma^2 + 4(\gamma_{i,t,s,l}^{(M-1)})^2}{\sigma^2 + 4(\gamma^*)^2}} \exp\left\{ -\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2} \cdot \left(\frac{1}{\sigma^2 + 4(\gamma^*)^2} - \frac{1}{\sigma^2 + 4(\gamma_{i,t,s,l}^{(M-1)})^2} \right) \right\}, \tag{C.4}$$

and M indicates the M-th iteration step in MCMC.

Simplifications to the Steps When ϕ is Known

Thanks to collapsing θ we will have to re-define $\psi_{i,t,s,l}$ as follows,

$$\psi_{i,t,s,l}^{-1} = 4\gamma_{i,t,s,l}^2 + \sigma^2 + \Delta_{i,t}/\phi.$$

Since $\boldsymbol{\theta}$ does not exist anymore, *Steps* 1 through 10 may not make sense. We would like to make least changes in the algorithm defined for general case. Let us re-define θ as follows:

$$\theta_{i,t} = X_i(t,:)\boldsymbol{\alpha}_i \tag{C.5}$$

With these new definitions in mind, MCMC steps will remain the same for *Step 1*, *Step 4* through *Step 8*.

Few changes are necessary in the following cases. In *Step 2* C.1 will be replaced by C.5. In *Step 9* ϕ will not be simulated like before. Instead ϕ will be assigned the known

value. In *Step 10* C.3 is replaced by the following,

$$\pi(\gamma_{i,t,s,l}|.) \propto \sqrt{\frac{1}{\sigma^2 + 4\gamma_{i,t,s,l}^2 + \Delta_{i,t}/\phi}} \exp\left\{ -\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2(\sigma^2 + 4\gamma_{i,t,s,l}^2 + \Delta_{i,t}/\phi)} \right\} p(\gamma_{i,t,s,l}).$$

and C.4 is replaced by the following,

$$\begin{aligned} LR &= \sqrt{\frac{\sigma^2 + \Delta_{i,t}/\phi + 4(\gamma_{i,t,s,l}^{(M-1)})^2}{\sigma^2 + \Delta_{i,t}/\phi + 4(\gamma^*)^2}} \exp\left\{ -\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2} \right. \\ &\cdot \left. \left(\frac{1}{\sigma^2 + \Delta_{i,t}/\phi + 4(\gamma^*)^2} - \frac{1}{\sigma^2 + \Delta_{i,t}/\phi + 4(\gamma_{i,t,s,l}^{(M-1)})^2} \right) \right\} .. \end{aligned}$$

In *Step 3* C.2 will be replaced by the following

$$\begin{aligned} \pi(\boldsymbol{\alpha}_i \mid .) &\propto \exp^{-(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^m)' P_i (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^m)/2} \prod_{j \geq 2}^{m} \mathbf{1}(\alpha_{i,j} \geq \alpha_{i,j-1}) \\ P_i &= \sum_t [X_i(t,:)' X_i(t,:)] \psi_{i,t,.,.} + \omega_i K^\delta, \quad \boldsymbol{\alpha}_i^m = P_i^{-1} \sum_t [X_i(t,:)' (\psi Z)_{i,t,.,.}] \end{aligned}$$

Posterior Density When ϕ is Known

$$\pi(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma} \mid \mathbf{X})$$

$$\propto \{\prod_{i=1}^{n} p(\tau_i)p(\delta_i)p(\boldsymbol{\alpha}_i|\omega_i)p(\omega_i|a,b)\}\{\prod_{i=1}^{n}\prod_{h=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}} p(\gamma_{i,t,s,l})\}$$

$$\times \{\prod_{i=1}^{n}\prod_{h=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}} [I(Y_{i,t,s,l} > 0)I(X_{i,t,s,l} = 1) + I(Y_{i,t,s,l} <= 0)I(X_{i,t,s,l} = 0)]$$

$$\sqrt{\frac{\psi_{i,t,s,l}}{2\pi}} \exp(-\frac{\psi_{i,t,s,l}}{2}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2)I(\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s})\}$$

$$\times \{\prod_{i=1}^{n}\prod_{h=1}^{T_i} \tau_i^{\frac{S_{i,t}-1}{2}} \exp(\frac{\tau_i \eta_{i,t}^{*\prime} \Sigma_{i,t}^{-1} \eta_{i,t}^{*\prime}}{2})\} \times \{\prod_{i=1}^{n}\prod_{h=1}^{T_i} \delta_i^{\frac{1}{2}} \exp(\frac{-\delta_i \varphi_{i,t}^2}{2})\}$$

$$\times \{\prod_{i=1}^{n}\prod_{t=1}^{T_i} \delta_i^{\frac{1}{2}} \exp(\frac{-\delta_i \varphi_{i,t}^2}{2})\} \times \{\prod_{i=1}^{n}\prod_{h=1}^{T_i} I(\theta_{i,t} = X_i(t,:)\boldsymbol{\alpha}_i)\}$$

Bibliography

Albers, W., Does, R., Imbos, T., and Janssen, M. (1989), "A stochastic growth model applied to repeated tests of academic knowledge," *Psychometrika*, 54, 451–466.

Albert, J. H. (1992), "Bayesian estimation of normal ogive item response curves using Gibbs sampling," *Journal of Educational and Behavioral Statistics*, 17, 251–269.

Andersen, E. B. (1985), "Estimating latent correlations between repeated testings," *Psychometrika*, 50, 3–16.

Andrews, D. F. and Mallows, C. L. (1974), "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102.

Andrich, D. and Kreiner, S. (2010), "Quantifying Response Dependence Between Two Dichotomous Items Using the Rasch Model," *Applied Psychological Measurement*, 34, 181–192.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011), "Assessment of school performance through a multilevel latent Markov Rasch model," *Journal of Educational and Behavioral Statistics*, 36, 491–522.

Berger, J. O. (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385–402.

Bock, R. D. and Aitkin, M. (1981), "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, 46, 443–459.

Bollen, K. A. and Curran, P. J. (2004), "Autoregressive Latent Trajectory (ALT) Models A Synthesis of Two Traditions," *Sociological Methods and Research*, 32, 336–383.

Bradlow, E., Wainer, H., and Wang, X. (1999), "A Bayesian random effects model for testlets," *Psychometrika*, 64, 153–168.

Brezger, A. and Lang, S. (2006), "Generalized structured additive regression based on Bayesian P-splines," *Computational Statistics & Data Analysis*, 50, 967 – 991.

Brezger, A. and Steiner, W. J. (2008), "Monotonic Regression Based on Bayesian P–Splines," *Journal of Business & Economic Statistics*, 26, 90–104.

Cai, L. (2010), "A Two-Tier Full-Information Item Factor Analysis Model with Applications," *Psychometrika*, 75, 581–612.

Chen, W.-H. and Thissen, D. (1997), "Local Dependence Indexes for Item Pairs Using Item Response Theory," *Journal of Educational and Behavioral Statistics*, 22, 265–289.

Chib, S. and Greenberg, E. (1995), "Understanding the metropolis-hastings algorithm," *The american statistician*, 49, 327–335.

Curry, H. B. and Schoenberg, I. J. (1988), *On Pólya Frequency Functions IV: The Fundamental Spline Functions and their Limits*, pp. 347–383, Birkhäuser Boston, Boston, MA.

Darrell Bock, R. (1972), "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," *Psychometrika*, 37, 29–51.

Davier, M., Xu, X., and Carstensen, C. H. (2011), "Measuring Growth in a Longitudinal Large-Scale Assessment with a General Latent Variable Model," *Psychometrika*, 76, 318–336.

De Boeck, P. (2008), "Random Item IRT Models," *Psychometrika*, 73, 533–559.

De Boor, C. (2001), *A Practical Guide to Splines (rev. ed.)*, New York: Springer.

D'Urso, P. (2000), "Dissimilarity measures for time trajectories," *Journal of the Italian Statistical Society*, 9, 53–83.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–102.

Embretson, S. (1991), "A multidimensional latent trait model for measuring learning and change," *Psychometrika*, 56, 495–515.

Everitt, B. S. (1981), *Finite mixture distributions*, Wiley Online Library.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011), "Cluster analysis," *SSRC Reviews of Current Research*, 11.

Ferrando, P. J. and Lorenzo-Seva, U. (2007), "An Item Response Theory Model for Incorporating Response Time Data in Binary Personality Items," *Applied Psychological Measurement*, 31, 525–543.

Fox, J.-P. and Glas, C. A. (2001), "Bayesian estimation of a multilevel IRT model using Gibbs sampling," *Psychometrika*, 66, 271–288.

Gaviria, J.-l. (2005), "Increase in precision when estimating parameters in computer assisted testing using response time," *Quality and Quantity*, 39, 45–69.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Glas, C. A. and Falcón, J. C. S. (2003), "A comparison of item-fit statistics for the three-parameter logistic model," *Applied Psychological Measurement*, 27, 87–106.

Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 100–108.

Heggeseth, B. C. (2013), "Longitudinal Cluster Analysis with Applications to Growth Trajectories," Ph.D. thesis, University of California, Berkeley.

Hsieh, C.-A., von Eye, A., Maier, K., Hsieh, H.-J., and Chen, S.-H. (2013), "A Unified Latent Growth Curve Model," *Structural Equation Modeling: A Multidisciplinary Journal*, 20, 592–615.

Jannarone, R. (1986), "Conjunctive item response theory kernels," *Psychometrika*, 51, 357–373.

Jeffreys, H. (1998), *The Theory of Probability*, Oxford Classic Texts in the Physical Sciences, OUP Oxford.

Johnson, C. and Raudenbush, S. W. (2006), *A Repeated Measures, Multilevel Rasch Model With Application to Self-Reported Criminal Behavior*, pp. 131–164, New York: Routledge.

Johnson, V. E. (2004), "A Bayesian χ^2 test for goodness-of-fit," *Ann. Statist.*, 32, 2361–2384.

Kaufman, L. and Rousseeuw, P. J. (2008), *Partitioning Around Medoids (Program PAM)*, pp. 68–125, John Wiley & Sons, Inc.

Klein Entink, R. H. (2009), "Statistical models for responses and response times," Ph.D. thesis, University of Twente.

Lindley, D. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 1, Probability*, Introduction to Probability and Statistics from a Bayesian View Point, Cambridge University Press.

Liu, Y. and Maydeu-Olivares, A. (2013), "Local dependence diagnostics in IRT modeling of binary data," *Educational and Psychological Measurement*, 73, 254–274.

Loeys, T., Rosseel, Y., and Baten, K. (2011), "A joint modeling approach for reaction time and accuracy in psycholinguistic experiments," *Psychometrika*, 76, 487–503.

Lord, F. (1953), "The Relation of Test Score to the Trait Underlying the Test," *Educational Psychology Measurement*, 13, 517–548.

Luce, R. D. (1986), *Response Times: Their Role in Inferring Elementary Mental Organization3*, vol. 8, Oxford University Press.

Martin, A. D. and Quinn, K. M. (2002), "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999," *Political Analysis*, 10, 134–153.

McCoy, S. I. (2010), "A trajectory analysis of alcohol and marijuana use among Latino in San Francisco, California," *Journal of Adolescent Health*, 47.6, 564–574.

Millar, R. B. (2009), "Comparison of Hierarchical Bayesian Models for Overdispersed Count Data using DIC and Bayes' Factors," *Biometrics*, 65, 962–969.

Milligan, G. W. and Cooper, M. C. (1985), "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, 50, 159–179.

Park, J. H. (2011), *Modeling Preference Changes via a Hidden Markov Item Response Theory Model.*, pp. 479–491, Boca Raton: CRC Press.

Patz, R. J. and Junker, B. W. (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models," *Journal of Educational and Behavioral Statistics*, 24, 146–178.

Prochazkova, J. (2005), "Derivative of B-Spline function," in *Proceedings of the 25th Conference on Geometry and Computer Graphics. Prague, Czech Republic.*

Ranger, J. and Kuhn, J.-T. (2012), "Improving Item Response Theory Model Calibration by Considering Response Times in Psychological Tests," *Applied Psychological Measurement*, 36, 214–231.

Rasch, G. (1961), "On General Laws and the Meaning of Measurement in Psychology," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine*, pp. 321–333, Berkeley, Calif., University of California Press.

Robert, C. P. (1995), "Simulation of truncated normal variables," *Statistics and computing*, 5, 121–125.

Robert K. Tsutakawa, M. J. S. (1988), "Approximation for Bayesian Ability Estimation," *Journal of Educational Statistics*, 13, 117–130.

Roskam, E. E. (1997), "Models for speed and time-limit tests," in *Handbook of modern item response theory*, pp. 187–208, Springer.

Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores," *Psychometrika monograph supplement*.

Sinharay, S., Johnson, M. S., and Williamson, D. M. (2003), "Calibrating Item Families and Summarizing the Results Using Family Expected Response Functions," *Journal of Educational and Behavioral Statistics*, 28, 295–313.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Tanner, M. A. and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, 82, 528–540.

Tarpey, T. (2012), "Linear transformations and the k-means clustering algorithm," *The American Statistician*.

Thissen, D. (1983), "Timed testing: An approach using item response theory," in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, ed. D. J. Weiss, pp. 179–203, Academic Press, New York.

Van der Linden, W. J. (2007), "A hierarchical framework for modeling speed and accuracy on test items," *Psychometrika*, 72, 287–308.

Van der Linden, W. J., Klein Entink, R. H., and Fox, J.-P. (2010), "IRT Parameter Estimation With Response Times as Collateral Information," *Applied Psychological Measurement*, 34, 327–347.

Verhagen, J. and Fox, J.-P. (2012), "Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses," *Statistics in Medicine*, pp. 2988–3005.

Wang, T. (2006), "A Model for the Joint Distribution of Item Response and Response Time using a One-Parameter Weibull Distribution," Tech. rep., CASMA Research Report.

Wang, T. and Hanson, B. A. (2005), "Development and Calibration of an Item Response Model That Incorporates Response Time," *Applied Psychological Measurement*, 29, 323–339.

Wang, T. and Zhang, J. (2006), "Optimal partitioning of testing time: theoretical properties and practical implications," *Psychometrika*, 71, 105–120.

Wang, X., Berger, J. O., and Burdick, D. S. (2013), "Bayesian analysis of dynamic item response models in educational testing," *The Annals of Applied Statistics*, 7, 126–153.

West, M. and Harrison, J. (1997), *Bayesian forecasting and dynamic models*, Springer.

Yao, H., Kim, S., Chen, M.-H., Ibrahim, J. G., Shah, A. K., and Lin, J. (2015), "Bayesian Inference for Multivariate Meta-Regression With a Partially Observed Within-Study Sample Covariance Matrix," *Journal of the American Statistical Association*, 110, 528–544, PMID: 26257452.

Yen, W. M. (1984), "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model," *Applied Psychological Measurement*, 8, 125–145.

Zellner, A. (1971), *An introduction to Bayesian inference in econometrics*, Wiley series in probability and mathematical statistics: Applied probability and statistics, J. Wiley.

Zerbe, G. O. (1979), "A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data." *Growth*, 43, 263–272.